DOCTOR OF PHILOSOPHY

The Effectiveness of Self-Assessment and its Viability in the Electronic Medium

Haig, D. Alexander J.

Award date:
2013

DOCTOR OF PHILOSOPHY

The Effectiveness of Self-Assessment and its Viability in the Electronic Medium

D. Alexander J. Haig

2013

University of Dundee

The Effectiveness of Self-Assessment and its Viability in the Electronic Medium

D. Alexander J. Haig

A thesis submitted in
fulfilment of
the requirements for
the degree of
Doctor of Philosophy

September 2013

Contents

Figures

Tables

Declaration

I, Alex Haig, am the author of this thesis. All references cited have been consulted, unless otherwise stated. It has not been previously accepted for a higher degree. The two systematic reviews (Chapters Two and Three) were joint research, the nature of which is detailed in each. The remainder is original research.

_____

Abbreviations and *Glossary*

ARCP Annual Review of Competence and Progression

BEME Best Evidence Medical Education Collaboration
www.bemecollaboration.org/

CASP Critical Appraisal Skills Programme
www.casp-uk.net/

CBD Case Based Discussion

CMT Core Medical Training

CI Confidence Interval
– an interval used to indicate the reliability of an estimated value

COP Certificate of Performance

CPD Continuing Professional Development

Deanery *- a regional organisation, within the structure of the UK National Health Service (NHS), responsible for postgraduate medical and dental training. (a)*

df degrees of freedom
– number of values in the final calculation of a statistic that are free to vary

DOPS Directly Observed Procedural Skill

DOTS Doctors Online Training System (2004-12)

ePortfolio *– an e-portfolio system for NHS staff groups; also known as NES ePortfolio*
www.nhseportfolios.org

e-portfolio *- a collection of electronic evidence assembled and managed by a user, usually on the Web. Such electronic evidence may include inputted text, electronic files, images, multimedia, blog entries, and hyperlinks. E-portfolios are both demonstrations of the user's abilities and platforms for self-expression, and, if they are online, they can be maintained dynamically over time. Some e-portfolio applications permit varying degrees of audience access, so the same portfolio might be used for multiple purposes. (a)*

FE Further Education

Foundation Programme
– a two-year structured programme of workplace-based learning for junior doctors that forms a bridge between medical school and specialty training. The programme aims to provide a safe, well-supervised environment for doctors to put into practice what they learned in medical school. (a)

Foundation School
- a conceptual group of institutions bringing together medical schools, the local deanery, trusts (acute, mental health and PCTs) and other organisations such as hospices. They aim to offer training to foundation doctors in a range of different settings and clinical environments. The

	schools are administered by central local staff, supported by the deanery. (b)
GMC	General Medical Council www.gmc-uk.org/
GPA	Grade Point Average (US)
ITE	In-Training Examination
JRCPTB	Joint Royal Colleges of Physicians Training Board www.jrcptb.org.uk/
Learn	*– replacement e-learning system for DOTS (see above) integrated in ePortfolio from 2012*
LMS	Learning Management System
MCAT	Medical College Admission Test
MCQ	Multiple Choice Question(s)
Mini-Cex	Mini-Clinical Evaluation Exercise
miniPAT	Mini Peer Assessment Tool
MMC	Modernising Medical Careers
MSF	Multi-Source Feedback
NES	NHS Education for Scotland www.nes.scot.nhs.uk/
NHS	National Health Service
NVivo	*– software for quantitative analysis* www.qsrinternational.com/products_nvivo.aspx
OSATS	Objective Structured Assessment of Technical Skills
OSCE	Objective Structured Clinical Examination
OR	Odds Ratio
	– a measure of association between an exposure and an outcome
PDA	Personal Digital Assistant (term largely superseded by "smartphone")
PDP	Personal Development Plan
PMETB	Postgraduate Medical Education Training Board (merged with GMC 1st April 2010)
PRHO	Pre-registration house officer
Quartiles	*– in this thesis, the four population groups defined by their self-assessment behaviour as defined by Kruger and Dunning*
QUESTS	*– critical appraisal criteria proposed by the BEME Collaboration (BEME Guide 1 www.bemecollaboration.org/BEME+Guides/)*
r	*– correlation coefficient*
Revalidation	*– the process by which a doctor demonstrates they can continue to practise, regulated by the GMC*
Role (ePortfolio)	*– group of users within the ePortfolio who share a defined number of system permissions and functions (e.g. trainee, educational supervisor)*
SA	self assessment
SD	Standard Deviation
	– a measure of the variation from the mean
SEA	Significant Event Analysis
SHO	Senior House Officer (note: term now obsolete)

SLE	Supervised Learning Event
SP	Standardised Patient
SpR	Specialist Registrar
SPSS	*– statistical software package* www-01.ibm.com/software/uk/analytics/spss/
SQL	*– a special-purpose programming language designed for managing data in relational database management systems (a)*
TAB	Team Assessment of Behaviour
Web 2.0	*– web sites that use technology beyond the static pages of earlier web sites, it does not refer to an update to any technical specification, but rather to cumulative changes in the ways software developers and end-users use the Web. A Web 2.0 site may allow users to interact and collaborate with each other in a social media dialogue as creators of user-generated content in a virtual community, in contrast to websites where people are limited to the passive viewing of content. (a)*

(a) from Wikipedia http://en.wikipedia.org/wiki/ (15/12/12)
(b) from UKFPO website www.foundationprogramme.nhs.uk/ (15/12/12)

Abstract

Background: Self-assessment is widely used across the health professions for a variety of purposes, including appraisal, CPD and revalidation. Despite numerous reported short-comings, the use of self-assessment is increasing, frequently on the requirements of regulatory bodies. Traditionally it has been a paper exercise, but in recent years self-assessment has appeared in electronic portfolios – a medium often used to collate assessments and other educational requirements. This thesis evaluates the effectiveness of self-assessment, in particular delivered via an e-portfolio, to determine if it:

- Improves the accuracy of perception of learning needs
- Promotes appropriate change in learner activity
- Improves clinical practice

Methods: This thesis is comprised of two systematic reviews and a case study. The first of two systematic reviews examines the evidence for effectiveness of self-assessment in the three research questions. The second evaluates the effectiveness of portfolios as a medium for postgraduate healthcare education. Both reviews are notable in that they employ systematic review methodology on non-clinical questions and amalgamate quantitative and qualitative data.

The final research component is an exploratory case study that tests the questions against a large data set (an entire training year of Scottish Foundation doctors) collated by the NHS ePortfolio. The case study provided the opportunity to separate groups of self-assessors identified by the literature, and compare the groups' self-scores against those of their supervisors and peers in the first and final post rotations; additionally, the groups' behaviour was matched against the literature for related educational activities recorded by the ePortfolio such as personal development planning. The case study also allowed the medium of e-portfolios to be itself evaluated in practice as an educational infrastructure. Through the comprehensive and iterative examination of the large dataset it became apparent that quantitative analysis was of limited value and qualitative analysis of elicited the richness on the data in context.

Results: With both reviews, the original research questions were unable to be fully

answered due to the paucity of evidence of sufficient quality; however, both did discover relevant related evidence. The self-assessment review found competent practitioners are the best able to self-assess whilst the least competent are the least able to self-assess. Peer assessment was found to be more accurate than self and better aligns with faculty/supervisor assessment. Feedback and benchmarking can improve self-assessment accuracy, especially for the most competent, and video can be seen to enhance this. There is no conclusive evidence that gender or culture effect self-assessment ability. Practical skills are better self-assessed than knowledge-based or "soft" skills.

The portfolio review found summative assessment reliability improved with multiple raters and discussion between the raters. Evidence on whether portfolio use aided reflection was mixed, possibly because it was dependent on individual conditions. The engagement and support of supervisors is key to portfolios being used properly, and there is some evidence portfolio learners are less passive then non-users. The time required to effectively use a portfolio is rarely considered.

Although many of the literature's findings were born out by the case study, the data also revealed (often by omission) many flaws in the use of self-assessment and related activities, many of which can be ascribed to the training year examined Much of the qualitative examination of text corresponded with the wider literature with low self-raters being over-critical of their often superior skills and high self-raters being over confident. However, there was some dissonance with the literature in the final component in that supervisor scoring conflicted with expectations whilst there text comments continued to match the literature.

Conclusions: Assessment in post-graduate health care is high stakes and resource-intensive. Self-assessment, and its use within an electronic portfolio, is demonstrated to have enormous potential if properly implemented.

1 INTRODUCTION

Self-assessment is increasingly being used and promoted for a variety of purposes across the health professions, including formative and summative assessment, identification of learning needs and quality assurance of education and training. The proliferation of self-assessment tools and processes continues, as it is advocated as a core component in maintaining professionalism and supporting life-long learning. Numerous national regulatory bodies in medicine and nursing include self-assessment in appraisal as well as professional monitoring and development, and increasingly it features as a key component in electronic portfolios. Despite the widespread and growing use, there is little evidence that self-assessment is effective in the scenarios it is being used.

1.1 EVIDENCE BASED EDUCATION: BACKGROUND

The term “evidence based medicine” came into use in 1992 and is commonly defined as “using the current evidence in the medical literature to provide the best possible care to patients”. Evidence-based medicine is based on the conceptual work of Archie Cochrane in the 1970s, and the methodologies developed by the McMaster Group lead by David Sackett and Gordon Guyatt in the 1990s.

Soon after, there was growing interest in extending evidence based practice/medicine to medical (and health) education. An inconsistency is noted by Van der Vleuten in Advances in Physiology Education (1995) that although clinical and biomedical researchers shared attitudes and approaches, “the academic attitudes of the researcher appeared to change when educational issues were discussed. Critical appraisal and scientific scrutiny were suddenly replaced by personal experiences and beliefs, and sometimes by traditional values and dogmas”.

A BMJ editorial in May of 1999 notes that although at least one billion pounds a year is spent on medical education there is a paucity of evidence to support it: “the funds available for research and development of medical education are tiny, amounting in total to little more than a couple decent grants in molecular biology” (Petersen, 1999).

The article goes on to detail the differences that preclude an easy transposition of evidence-based medicine to evidence based medical education. These include the (lack of) uptake of medical education theories and publications by a wide audience and the perception that educational research and practitioners can be inward-looking. A second substantial point is that study designs employed are in the large majority qualitative, with extremely few randomised controlled trials (which arguably are not an appropriate or practical design for educational interventions).

1.1.1 Best Evidence Medical Education Collaboration

1999 saw the formation of Best Evidence Medical Education (BEME) which sought to introduce evidence based practice to medical (and health) education. Citing decision-making in the discipline as often subject to the forces of political, professional and public demand rather than any objective evidence, BEME set out to not only produce systematic reviews for medical education but to gradually shift the professional culture from opinion-based to evidence based. From its inception the BEME movement produced a prescriptive methodology (QUESTS) (Harden, 2000) to aid the researcher and practitioner.

Systematic reviews combined and/or synthesise all the best evidence available to answer research questions and inform best practice. Although their use in the clinical arena is well established, their potential value to other areas (sych as Education, Social Welfare, International Development) only came to be examined a decade and a half ago. BEME focused on medical (and later health) education to empower policy-makers and individuals with the ability to base their professional decisions on comprehensive analysis (and when possible synthesis) of all relevant research findings.

The BEME website currently (10.03.2013) lists twenty completed systematic reviews, with another seventeen in production. However, none of these adhere to the originally proposed methodology and each has employed methods of its own, often very different from other reviews under the BEME banner. Educational research nearly always produces evidence that is too heterogeneous for quantitative synthesis, so opportunities to use a Cochrane Review type model will be very rare (Clegg 2005;

Dixon-Woods 2006). Similarly, educational research methods themselves are wide ranging and frequently adapted. There are common aspects across (many) BEME reviews; for example, most use Kirkpatrick's hierarchy to gauge an intervention's impact on the participants, though most reviews modify the base framework to meet their individual requirements.

The problems encountered by the BEME review groups mirrored those faced in educational research itself, namely the difficulty in retrieving evidence, the quality of the studies and the challenges of employing meticulous methods in the complex collective relationships that comprise educational settings (Dauphinee, 2004). It is therefore not surprising that the methodology that underpins evidence based medicine cannot easily be transposed to an educational setting; however, despite this the drive towards demonstrating evidence to support decision making underpins professionalism, and making educational research more transparent and objective remains a worthy goal.

This thesis considers the effectiveness of two broad subjects in medical education: the use of self-assessment and the use of portfolios. Both had an extensive evidence base that had no recent or comprehensive synthesis.

1.2 OUTLINE

This thesis builds upon multiple projects examining the evidence for the effectiveness of professional self-assessment in health care and how electronic portfolios, can support and enable self-assessment. It is comprised of three major research projects: two systematic reviews and a large case study. Whilst the two systematic reviews focus on the effectiveness of self-assessment and portfolios in turn, they also heavily informed the design of the case study which uses the NHS ePortfolio as the tool to test the results of the first review in a "natural laboratory" of a year's postgraduate medical training. The timeline for this thesis' component parts is shown in Figure 1.

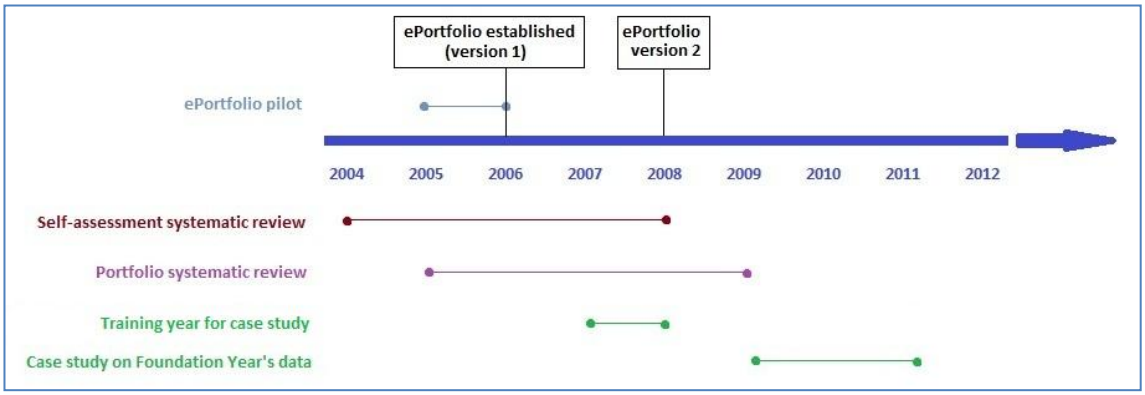

Figure 1. Thesis Timeline

This thesis is driven by three central questions on self-assessment, namely whether there are self-assessment interventions that:

1. **Improve the accuracy of learner perception of their learning needs?**
2. **Promote an appropriate change in learner activity?**
3. **Improve clinical practice?**

It examines medical trainees engaged with self-rating/assessment processes and how these relate to their wider education. These three questions were core to the first systematic review and were later re-examined against a full year's training data in the case study within the ePortfolio. The second systematic review examined portfolios as a medium and informed both the case study, as well as subsequent development of the ePortfolio system itself.

Whilst systematic reviews are widely used to support evidence-based clinical practice, their use in educational interventions is comparatively new. However, the methodological challenges of applying this study design in education are significant, and the benefits of utilising the most robust methods were in both cases judged to outweigh the challenges. Importantly, in both cases, the research questions had never been addressed by comprehensive evidence synthesis.

The first systematic review (Colthart, et. al. 2008) attempted to answer four central questions (the fourth relating to improvements in patient outcomes was subsequently dropped) relating to the effectiveness of self-assessment in specific areas, but was

beset with numerous problems in the evidence base including the poor quality of studies and the lack of an external measure of self-assessment. Nevertheless, the review was able to confirm a number of assumptions held about self-assessment (that it is less accurate than peer assessment, and that poor performers are also the worst self-assessors), as well as identify specific gaps in the evidence base. These assumptions and gaps were then tested in depth within the case study and were ultimately important to the design of the research tool itself (the ePortfolio).

The second systematic review (Tochel et. al., 2009) encountered similar problems to the first as it attempted to determine the genuine outcomes of portfolios use and the confounding variables that underpin variation across populations. And despite an evidence base that was more heterogeneous than the first review, the portfolio review was able to establish effective factors within portfolio use, as well as describe the benefits and risks of using them on an electronic platform. The findings were integrated into the design of the case study research tool, as the ePortfolio operationalized self-assessment in a new medium, allowing the case study to determine not if e-portfolios worked, but whether they could enable self-assessment.

The primary research component of this thesis (Chapter 5) was to design a research methodology to test the assumptions and gaps identified by the self-assessment review, whilst drawing upon the portfolio reviews to examine the research tool, the NHS ePortfolio. The self-assessment review was instrumental in identifying appropriate research groups for the case study – quartiles of self-raters within an entire year of Scottish postgraduate medical trainees. Multi-source feedback scores within the ePortfolio allowed the comparison of self-ratings with ratings of peers and supervisors, as well as the other assessment and educational tools and processes supported by the trainee doctors' ePortfolios. This work used both quantitative and qualitative evidence to give a more comprehensive understanding of the issues in the theoretical literature. It is not a validity test of assessment data, but a detailed examination of individuals' journeys in a given training year after demarcation by relevant self-assessment level and correlated with external ratings.

The use of self-assessment continues to expand internationally across healthcare: tens

of thousands of UK doctors and dentists use e-portfolios in training, and increasingly with revalidation. It is crucial for both the professions and the public to have their use supported with evidence of their effectiveness. This project's discussion focuses on the potential of e-portfolios to improve the accuracy and value of self-assessment, improve learner awareness of standards and provide timely and rich feedback.

2 SELF-ASSESSMENT SYSTEMATIC REVIEW

2.1 REVIEW OF EVIDENCE

Self-assessment has come to be seen as a fundamental component of learning across the health professions, core to appraisal systems and developing clinical learning, as well as a cornerstone of professional behaviour, often with an inherent assumption that learners can and will identify their own learning needs (Gordon, 1991).

But despite the increasingly high profile given to self-assessment, there had been no analysis of the evidence on the topic since a narrative review was published in 1991 (Gordon *et al*) that determined that "self-assessment skills remain undeveloped during training". His paper identified a tendency towards over-confidence in self-assessment, particularly for those with the lowest levels of ability. A second finding was that clinical skills, as opposed to knowledge or communication skills, showed higher correlation coefficients between self and external raters. Global self-assessments were also seen to have significant impact on individual self-assessments – perhaps as closely related to an individual's perception of themselves as is their previous performance. Finally, there was some evidence that self-assessment could be made more accurate with greater involvement with students in the learning, clarity of measure and feedback and resolution of external raters with self.

This systematic review sought to comprehensively retrieve and synthesise all good quality evidence published since the Gordon paper, regardless of study design. Gordon's 1991 review, as well as subsequent influential papers (Ward *et al* 2002), described that it was largely quantitative papers that indicated poor accuracy of self-assessment. However, these studies often used non-validated scales or external measures and were thought by themselves to not necessarily give a comprehensive analysis of the effectiveness and precision of self-assessment.

The definition of self-assessment was been problematic in itself. Gordon referred to "valid self assessment" as "judging one's performance against appropriate criteria" while defining "accurate self assessment" as "gaining reasonable concurrence between self-claimed and other, validated measures of performance". For Ward, self-

assessment is the “ability to accurately assess one’s own strengths and weaknesses” and like Gordon, sees the ability as being “critical to lifelong learning”. Eva and Regehr (2005) purport that the complexity of self-assessment means it does not lend itself to simple or concise definition and instead advocate professionals continually relating to incidents and self-assessing on these individual strengths and weaknesses.

As self-assessment involves self-referential thinking there is an inherent overlap with the psychological literature, particularly the concept of self-efficacy. Within psychology, self-efficacy is commonly held to mean an individual’s belief in their own abilities to achieve certain goals. Bandura (1994) refers to self-efficacy as “people’s beliefs about their capabilities to produce designated levels of performance that exercise influence over events in their lives”, which distinguishes from self-assessment in that it can be seen as a strong influence over performance which can lead to a greater chance of success. This review therefore only included self-efficacy papers if they described a method or tool of self-assessment.

After considerable debate, this systematic review was established using a definition of self-assessment as, “a personal evaluation of one’s professional attributes and abilities against perceived norms”.

The review group was comprised of members of a variety of backgrounds, including medicine, nursing, information science and research methods, who were employed by both the NHS and academic institutions. The author’s role in the group was initially to raise awareness of the BEME Collaboration and garner interest in conducting an educational systematic review. From there, the author took a lead in establishing the team and working to establish the methodology within this group in conjunction with other BEME groups. The membership was assembled before self-assessment was determined to be the subject area, although it had been favoured by most members as the priority area of interest from the onset. The paper (Colthart et al 2008), published in Medical Teacher in 2008 was the culmination of several years’ work. Why the review did not find evidence of sufficient quality and quantity to answer its initial questions, it did identify many factors that influence self-assessment, as well as areas that were need of urgent research.

2.2 SELF-ASSESSMENT: PREVIOUS RESEARCH

An influential review of the accuracy and reliability of self-assessment in healthcare settings was published by Gordon *et al* in 1991. In this paper the authors characterised four types of study within the subject area.

The first type of study, "Experiments in which self-claimed factual knowledge was tested against verifiable facts", revealed an inclination amongst learners to over-estimate their abilities, especially when their knowledge of the subject area was lower. The next type of study, "Studies in which health professions' trainees viewed samples of their own clinical behaviour on videotape and assessed their performance using behavioural rating instruments" compared ratings of clinical skills between student (self) and faculty. This showed video-taped reviews yielded better self-assessment results, in particular when grading framework were recalibrated. The third type of study, "Global self-assessments of performance based on extended periods of supervised functioning in clinical training environments", had the authors concluding that "global self attributions" have a substantial impact on self-assessment – possibly as much impact as an individual's previous performance. The final type of study, "Studies of innovative training programmes in which valid and accurate self-assessment was an explicit goal and in which specific strategies for improving self-assessment skills were used", where external performance measures aided the accuracy of the students' self- assessment. Each of these (four) study types showed higher correlation coefficients amongst the more specific and clinical skills measured. The increased accuracy of self-assessment in measuring "hard" (clinical) over "soft" (e.g. communication) skills is the subject of consequent research and forms a part of this thesis. The review concluded that the skills required for accurate self-assessment within a healthcare training setting were not sufficiently developed, but also highlighted the lack of robust evidence on the subject.

The following year (1992) Gordon went on to review programmes of self-assessment and reported a similar scarcity of high quality evidence, as well as the fact that studies on self-assessment programmes did not reference previous work in the area. Nevertheless, the review of programmes did identify two characteristics common to

reliable and accurate self-assessment. The first, in common with the 1991 review, cited programmes with explicit and formal requirements to link learners' self-assessments with external measures. The second was the (unfounded) assumption across the programmes that learners would comprehensively collect and examine evidence on their performances.

Gordon's work highlighted the need for a further systematic evaluation of the consequent decade and a half of research in light of his findings. The confirmation or rejection of the findings, or the absence of sufficient evidence, would inform this thesis' case study (Chapter 5) which also had its construction informed by the results of the portfolio systematic review (Chapter 3).

Although outside of healthcare, a seminal paper published in 1999 by Kruger and Dunning had considerable impact across subsequent self -assessment studies, and deserves specific individual mention. Their water-shed paper examined whether the least competent individuals had greater difficulty assessing themselves and whether it was a lack of meta-cognitive skill that was responsible for this lack of ability.

Kruger and Dunning observed their research population in quartiles (ranked according to their self-assessment scores), and found the lowest performing quarter the least able to accurately self-assess and overestimated their abilities. In contrast, the top quartile, whilst being more accurate self-assessors than the lowest quartile, underestimated their abilities. After benchmarking, the top quartile revised their ratings upwards, making them more accurate; however, the lowest quartile also revised their own ratings upwards making them more inaccurate.

The best skilled individuals were seen to operate under a "false consensus effect" with the assumption that their peers would be of similar abilities, but could re-calibrate with benchmarking. The lowest quartile was unable to gain insight into their own lack of ability from peer observation, and indeed became slightly less insightful. Kruger and Dunning (1999) concluded with a final study that demonstrated the lowest quartile could improve their self-assessment ability – but only when given the opportunity to improve their meta-cognitive skills (i.e. the awareness or ability to analyse one's own thinking and learning processes) which allowed them to realise their deficiencies.

Although the authors conducted the research in a psychology setting (looking at skills such as logical reasoning, humour and grammatical structuring) the results of this paper have subsequently informed much further research in a wide variety of settings and have been replicated in healthcare (Edwards et al., 2003, Ehrlinger et al., 2003, Hodges et al., 2001, Lane et al., 2004, Mandel et al., 2005). The original paper was based on questionnaires and thematic analysis of text, yet many consequent healthcare studies that sought to replicate the findings used experimental design. This case study of this thesis was exploratory work and given the data set contained both quantitative and qualitative data set out to utilise both to answer its research questions.

Previous reviews (Gordon 1991; Ward *et al*. 2002) suggest that much of the evidence for poor accuracy of self-assessment was based on quantitative studies, some of which used group analyses to compare ratings of students and teachers, often with un-validated rating scales. Individual accuracy in identifying strengths and weaknesses would not be identified in such studies. These issues have been discussed at length by Ward *et al*. (2002) and will be explored in more detail later in section 2.8. The case study (Chapter 5) was designed to incorporate as much qualitative evidence as possible in an attempt to balance this deficiency.

For the reasons given above, it is unlikely that such studies would have given us a complete picture of the accuracy and usefulness of self-assessment in the health professions. In the review undertaken as part of this thesis, therefore, studies were not excluded based on particular research methods, but were selected on the basis of study quality and whether the conclusions were important and likely to be applicable in contexts other than that of the original research. As noted in the introduction, the importance of updating the understanding of self-assessment in clinical education is emphasised by the increasingly widespread assumption that learners will accurately identify their own learning needs through self-assessment.

Given that self-assessment is generally accepted by policy makers as a prerequisite for continuing professional development (CPD) in the health professions, the specific review question centred on the evidence around self-assessment interventions. In line

with other Best Evidence Medical Education (BEME) reviews (Dornan *et al*. 2006; Hammick *et al*. 2007) the review determined if there was evidence of self-assessment interventions improving outcomes at each level of Kirkpatrick's evaluation hierarchy (Section 2.4.5) (Kirkpatrick, 1967).

The role of the author within the group was composite. Initially the author drew the group together from colleagues interested in the subject area and/or working on a systematic review in education. The heterogeneous nature of the group's interests and experience was reflected in the agreed research aims (below). In addition to being key author, I designed, tested and conducted the systematic search, advised on critical appraisal and designed the electronic reviewing system in conjunction with a programmer.

2.3 Research Questions

There were four central *research questions* for the systematic review, that were developed iteratively by the review group from their own research interests and skills, and well as a preliminary examination of the literature and discussion. Each question asked whether there are effective self-assessment interventions which:

I. **Improve the accuracy of learner perception of their learning needs?**
II. **Promote an appropriate change in leaner activity?**
III. **Improve clinical practice?**
IV. **Improve patient outcomes?**

There were an additional two subsidiary research questions:

- *What are the factors affecting the accuracy of self-assessment in relation to other assessments such as peer and external?*
- *What are learners' and teachers' perceptions of, and attitudes to, self-assessment?*

2.4 Study Selection

The following section describes the objectives and methods used for this review and various aspects of the selected studies.

2.4.1 Objectives

2.4.2 Study Identification

This self-assessment systematic review had the following *objectives*:

- Identify the scope of the research on the effectiveness of self-assessment methods
- Review the evidence of the impact of self-assessment methods on:
    - Identification of learning needs
    - Learning activity
    - Clinical practice
- Identify the perceived value of self-assessment to learners
- Make recommendations for further research and practice

The inclusion and exclusion criteria were defined by the research question(s) to ensure all relevant papers were retrieved. The selection criteria were:

1. Is it about self-assessment?
2. Is it set in a clinical training context?
3. Does it have any one of the following:
    a. An evaluation of the self-assessment method or tool?
    b. Offer important information about attitudes towards/perceptions of self- assessment?
    c. Is it a comparison study (measuring accuracy of self-assessment against some other assessment)?
    d. Does it describe an impact of self-assessment on teachers and/or learners?

Studies were excluded if they were not primary research (e.g. reviews –these were included in the Discussion), no assessment of intervention and/or its impact, not in a clinical context, not self-assessment (audit), self-assessment used to evaluate another programme or intervention (blind tool) or there was no structured self-assessment method described.

BEME groups were and are expected to adapt and test the general guidance to their specific topic, and develop a transparent and objective system of peer review. The research protocol was submitted to BEME for peer review. Details follow below.

2.4.3 Types of Studies – Research Designs

All research designs were considered. These categories were derived from the initial review of abstracts and reflect the content of the abstracts rather than formal theoretical frameworks within educational research. Many studies were not explicit about their underlying theoretical framework, and the aim was to incorporate all relevant approaches.

Studies were included that compared the accuracy of self-assessment in a variety of clinical settings with peer or tutor assessment in order to determine if particular groups of learners are more accurate than others in self-assessment. Also considered were studies that explored the attitudes of learners and teachers to self-assessment. To help understand the range of methods employed within these research designs information was recorded on data collection methods (e.g. interviews, questionnaires, and observations) and analysis (qualitative, quantitative or both). The type of clinical setting in which the intervention took place was also recorded and the professional context involved. Finally, synonyms and definitions of self-assessment used by different authors were noted.

2.4.4 Self-Assessment Intervention Types

All forms of structured self-assessment which included an explicit intervention method or tool were considered. In addition studies of interventions to improve the

effectiveness of self-assessment were included.

2.4.5 Participants

All professions in clinical practice including chiropodists/podiatrists, complementary therapists, dentists, dieticians, doctors, hygienists, psychologists, psychotherapists, midwives, nurses, pharmacists, physiotherapists, occupational therapists, radiographers and speech therapists were included, as were clinical undergraduate students from these specialties.

2.4.6 Outcome Measures

Outcome measures were based on an extended version of Kirkpatrick's (1967) model of outcomes at four levels as shown in Figure 2 (BMJ, 1999). Also included were outcome measures of accuracy of self-assessment and the factors influencing self-assessment and additional predetermined and unintended outcomes were also accepted. The (adaptive) use of the Kirkpatrick model is not mandated for BEME groups, but it has been used or adapted by most groups as a useful framework for the evaluation of evidence related to learning.

Figure 2. Extended Version of Kirkpatrick's Model

2.4.7 Search Strategy

A comprehensive literature search was conducted across all sources relevant to professional education in a clinical context.

The database search covered all relevant health as well as educational databases, and included: Medline, CINAHL, BNI, Embase, EBM Collection, Psychlit, HMIC, ERIC, BEI, TIMElit and RDRB. The strategies were designed and tested for maximum sensitivity to ensure no potentially relevant papers were missed. The search limits were from January 1990 to February 2005 and did not limit by language, geography, or research methodology. An updating search was conducted in January 2006 to include evidence published during the course of this group's analysis.

The results of the database search were augmented by further methods. A cited reference search was conducted on the core papers of relevance examining which papers these cited, and in turn which future papers referred back to the core. This is a method BEME has found very effective for retrieving relevant papers that imperfect educational descriptors within clinical databases fail to adequately describe. Grey literature (evidence not formally or commercially published) searches were also conducted along BEME methodology (further information on grey literature searching is in section 3.3.1, as the second review's topic was much more likely to have this type of evidence).

Finally, hand searches were conducted across the most relevant journals: *Academic Medicine*, *Medical Teacher*, *Medical Education, Nurse Education in Practice* and *Education for Primary Care,* as it is recognized that electronic indexing of clinical education terms and clinical educational journals was unreliable at times throughout that period. Titles suggesting a focus on self-assessment that had not already been identified were obtained for examination of abstract and, if indicated, full text. References in full text articles were explored for additional relevant citations.

The original list of retrieved articles was visually scanned to determine whether they potentially fulfilled the research questions. From this list the abstracts were obtained. All abstracts were viewed by at least two group members to decide if a full text version of the article should be obtained. The process of the review is summarized in Figure 1,

which shows that 77 papers were agreed for retrieval in full; of these 39 were not considered as informative, 32 were, and an additional 6 papers were included for their relevance although they did not satisfy all the inclusion criteria (e.g. a review rather than primary research).

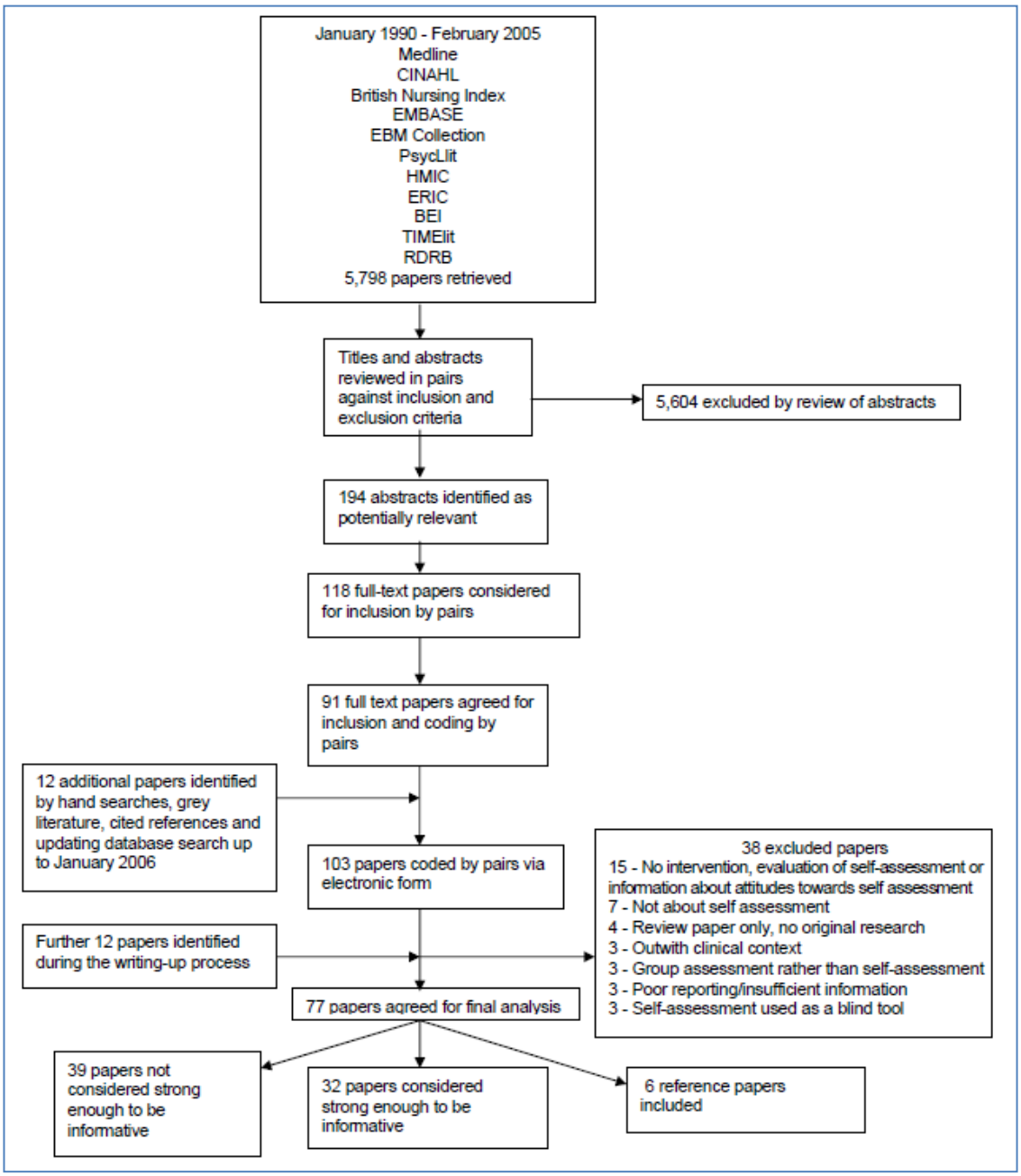

Figure 3. Flowchart of Search and Selection Strategy

2.4.8 Data Abstraction

A coding form was devised from the BEME standard version, containing sections to determine the strength and relevance of the study to the research questions, as well as the rigour of the study design itself. The latter sections were adapted from the NHS Critical Appraisal Skills Programme (CASP) tools, widely-used critical appraisal instruments created to objectively evaluate specific research methodologies, and were found within the Solutions for Public Health website.[1] In addition an instrument to assess the quality of comparative studies was devised by the group. The checklists appear in the coding sheet.

The coding sheets were designed to permit consistency across the different qualitative and quantitative approaches to data collection. All members of the review team independently coded a selection of papers into the data abstraction sheets to validate the coding sheets for utility and completeness.

All full papers were then read by two group members, using the final version of the coding sheet to extract and assess data. As the group was split between different sites across the United Kingdom (Edinburgh, Glasgow, Newcastle, Leeds and Birmingham), a web-based coding form was developed to enable geographically separated pairs to code and agree data. Papers which on full reading did not meet the inclusion requirements were rejected and the reasons recorded. Abstracted data included a detailed checklist for the different types of research method employed. Reviewers were asked to rate;

- the appropriateness of the design of the study to answer the research questions posed,
- how well the design was implemented,
- the appropriateness of the analysis,

and to comment on concerns. They were then asked to comment on what level of the Kirkpatrick Hierarchy (Kirkpatrick 1967) the outcomes related to using the adapted version (Table 2) tailored to the research objectives of this systematic review.

[1] www.sph.nhs.uk/

Additionally reviewers identified references cited in these papers that might be of interest to the review and where appropriate these were obtained.

Table 1.Group Scoring of Strength of Findings and Overall Importance

	Grade	Description
Strength of findings of the paper	**1**	No clear conclusions can be drawn. Not significant.
	2	Results ambiguous, but there appears to be a trend.
	3	Conclusions can probably be based on the results.
	4	Results are clear and very likely to be true.
	5	Results are unequivocal.
Overall importance of the paper	**1**	Papers with numerous deficiencies in the rigour or appropriateness of the methodology or the statistical analysis
	2	Papers with some deficiencies in the rigour or appropriateness of the methodology or the statistical analysis
	3	Papers with doubts about the rigour or appropriateness of the methodology or the statistical analysis
	4	Papers with rigorous methodology and appropriate statistical analysis, but doubts about adequate sample size
	5	Papers with generalisable findings, rigorous methodology, adequate sample size, and appropriate statistical analysis.

Following data extraction of each paper the two group members independently scored them on a scale of 1 to 5 for the strength of the findings (

Table 1). Papers where the conclusions were not supported by the evidence presented (i.e. grades 1 and 2) were not considered further as their quality was not considered for inclusion. The perceived overall importance of the paper in terms of the rigour with which it was conducted, relevance, and generalisability was also graded independently by both reviewers. Again papers with grades 1 and 2 were discarded. The reviewing pair then consulted and agreed final scores for the paper. As with the abstracts, any discrepancies were usually resolved through discussion between the pair. Inter-reviewer agreement was high, with adjudication being required on only three occasions.

Papers that scored 4 or above on either *strength of findings* or *importance* were considered to be higher quality papers and were included and reported fully in the review. All these papers were read again and summarized in an abbreviated format by three members of the team. 'Borderline' papers (rated 3 on strength of findings and on

importance) were also reviewed independently to ensure that no higher quality paper had been excluded.

2.4.9 Analytical Procedures–Synthesising the Findings

Although the author was prepared if possible to undertake meta-analysis, it was recognised that very few of the variables coded were likely to be ratio data, with some interval data. Most of the data were categorical and insufficiently homogeneous to allow meta-analysis of results. The review therefore was largely descriptive, with the results reported through a narrative framework that focused on key themes. These are summarized below and form the subheadings for reporting the results. The key themes were:

- Peer Assessment and faculty ratings
- Individual characteristics
- Gender
- Cultural differences
- Insight
- External factors
- The purpose of the self-assessment task
- Practical skills versus theoretical knowledge
- Factors influencing self-assessment
- Benchmarking
- Video and verbal feedback
- Instruction
- Experience
- Perceptions and attitudes towards self-assessment

Each member of the review team undertook to synthesise data from papers that were considered to be of higher quality for one or more of the themes.

2.5 Search Results

Despite very inclusive strategies being employed (5,798 total hits were recorded) the conventional strategies were unable to retrieve all papers within the databases searched. The search specificity (the percentage of the returns that were actually relevant to the topic) was particularly poor at 3.3% and therefore time consuming for the group as thousands of false hits had to be discarded. This was due to ambiguities around searching for clinical education literature already researched by BEME but also to the lack of clarity and consistency ascribed to the concept of self-assessment itself. Search sensitivity (the percentage of the total relevant papers retrieved) was average, at 91%.

Although the search did not limit by geography or language, two thirds of the final papers were North American and over four fifths came from English-speaking countries. Homogeneity was also evident with regards to study design; while this group considered all research methods, less than 5% of included papers used only qualitative methods.

Table 2. Kirkpatrick's Hierarchy Adapted to Self-Assessment

Level	Description
1 - Reaction	These cover learners' views on the self-assessment experiences, its perceived usefulness, possible general positive and negative effects on learning, self- esteem, relationship with tutors and peers.
2 - Modification of attitudes / perceptions	These outcomes relate to specific perceived changes in individuals in respect to their perceptions of knowledge and skill in the tested area, specific impact on personal self-esteem and relationships with tutors and peers.
3 - Change in learning behaviour	Recorded change in learning behaviour as a result of a self-assessment intervention.
4a - Behavioural change	Actual change in clinical practice as a result of a self-assessment exercise.
Level 4b - Change in patient outcomes	Any improvements in the health and well-being of patients/clients as a direct result of self-assessment intervention. Where possible objectively measured or self-reported patient/client outcomes will be used, such as: health status measures, disease incidence, duration or cure rates, mortality, complication rates, readmission rates, adherence rates, patient or family satisfaction, continuity of care.

2.5.1 Methodological Quality of Included Studies

In many assessed papers, conventional good research practice was either not followed or the report of the study did not allow the reader to critically evaluate the study, as key pieces of information were not reported. The review has identified a variety of such problems and these are outlined below.

- Assessment instruments used in some studies were either not validated or no reference was made to their reliability and validity.
- There was a frequent assumption that expert opinion provided a gold standard, yet it was rare for validity or reliability of the expert opinions to be examined.
- The use of group means in some comparison studies ignored individual variation in self-assessment ability.
- In some studies control groups were needed but not used.
- It was rare for power calculations to be provided. Few studies were set up to test specific hypotheses, and most were limited to correlational analyses.
- Sampling and selection strategies were not stated in many studies, which meant that assessments could not be made of how representative the study participants were of their populations. Likewise many studies failed to present data on non-participants, which casts doubt on the representativeness of the sample.
- Inadequate explanation of missing data.
- Statistical methods were unclear.
- Study conducted at a single institution bringing into question the generalisability of the study.
- No clear information presented on how qualitative data were analysed.

The extent of the problems was surprising, but was common with the other BEME Systematic Review Groups.

The aim of several papers was to correlate a self-assessed measure against an external measure. Typically the external measure was the judgement of an assessor (peer, faculty, tutor or clinical preceptor) or a criterion measure such as an examination or checklist. The validity and reliability of these external measures was rarely reported.

This section reports the specific research findings from the 32 papers which scored 4 or above on either *strength of findings* or *importance*, which were the criteria for a paper to be considered of high quality.

Results are presented firstly in terms of their ability to answer the original research questions for the review, and then themes which emerged from the papers. Each theme forms a subheading in Section 2.6 below.

2.5.2 Research Questions

Few papers treated self-assessment as an intervention in itself, and none of the high quality papers looked specifically for changes as a result of undertaking self-assessment alone.

Are there effective self-assessment interventions which:

1. *improve the accuracy of learner perception of their learning needs?*

The majority of the studies found addressed the accuracy of self-assessment compared with an external assessment, but none of the high quality studies attempted to either measure change in perceptions of learning needs, or to find a valid assessment of learning needs against which to compare self-assessed needs. Interventions to improve the accuracy of self-assessment are discussed in a separate section below.

One paper that was difficult to classify did address the assessment of learning needs in children's hospice doctors (Amery & Lapwood 2004). This study was felt not to meet the inclusion criteria as there was no external comparator nor was there an evaluation of the self-assessment method. The findings, however, were interesting in that they highlighted the different learning needs identified when doctors completed questionnaires, and when they had an interview based on incidents reported in an educational diary. The authors suggest that a variety of methods are needed to fully identify learning needs, with 'self-perception analysis' being needed in addition to facilitation and diary keeping to help identify the areas that subjects don't know that they don't know.

2. *promote appropriate change in learner learning activity – Kirkpatrick level 3.*

None of the high quality papers reported any self-assessment intervention that led to a

change in learner's learning activity.

3. *improve clinical practice/improve patient outcomes – Kirkpatrick level 4.*

Only two papers addressed this question: Ericson *et al*. (1997) was recorded on the database as providing evaluation at level 4. The self-assessment exercise was carried out on 41 dental students and was accompanied by clinical guidelines, so it could be that the main educational effect was related to students following the guidelines rather than being the result of self-assessment. There was good agreement between tutors' and students' ratings (the same rating was given in 87% of instances, 10% of students underrated themselves, and 3% over-rated). This study suggests that the use of guidelines might aid self-assessment, but there was no control group. It does not present any evidence that self-assessment on its own has any impact at any Kirkpatrick level.

The second paper recorded on the database as Kirkpatrick level 4 was Biernat *et al*. (2003). This study compared faculty assessments with residents' self-assessment skills of their performance in an interview with a simulated patient portraying dementia. Twelve residents undertook a videotaped interview then completed a checklist of behaviours carried out in the interview. The videotape was rated by a faculty member, then residents were able to review the tape with the programme director for feedback and additional instruction. The residents completed an evaluation form, all of them reporting that the self-assessment tool was useful (Kirkpatrick level 1). One comment indicated that the experience would change the way the resident treated patients with memory loss, and another reported being encouraged to improve knowledge (Kirkpatrick level 2). There was no test of whether the practice of the residents changed, or any measure of change in patient outcomes.

In summary, there were not any high quality papers found to answer the main research questions, based on Kirkpatrick's hierarchy. However, some useful evidence was found on the subsidiary research questions and on other themes relating to self-assessment. Section 2.6 below summarizes the findings under sub-headings which reflect these themes. To facilitate interpretation, the text under each sub-heading includes a summary discussion. It is hoped that this will help the reader, rather than

having all the comments in a separate discussion section, which would lead to repetition and difficulty in linking the findings with the relevant section of the discussion.

2.6 Themes Relating to Self-Assessment

The following section examines the themes the group discovered common to multiple papers. These themes will form the assumptions and gaps to be tested in the case study.

2.6.1 Peer Assessment and Faculty Ratings

A number of studies have specifically addressed the question of peer assessment in the context of self-assessment. Typically self-assessment was correlated against both peer ratings and expert opinion which may be represented by faculty or a tutor. The research suggests a consistent pattern of results in relation to how self-assessment rates against peer assessment. The following studies typify the general conclusion across a number of studies that individuals are more able to accurately assess their peers' ability than their own.

Rudy *et al*. (2001) compared self-assessment, peer and faculty evaluations of interviewing skills for 97 first year medical students. Although correlations were modest they found that individuals gave their peers a more balanced assessment in comparison to how they rated themselves. Correlations between self and peer ratings ($r=0.29$, df (degrees of freedom) $=89$, $p=0.008$) and between faculty and peer ratings ($r=0.50$, $df=86$, $p=0.0001$) were statistically significant. The correlation between self and faculty composite scores showed marginal statistical significance ($r=0.19$, $df=80$, $p=0.08$). This leads them to conclude that students are capable of assessing their peers but have difficulty in accurately evaluating their own performance. Sullivan *et al*. (1999) used a similar methodology by comparing self, peer and faculty ratings in the setting of a problem based tutorial group for 154 third year medical students.

They found that the medical students were not able to identify their own strengths

and weaknesses as compared to their peers and faculty. Three areas were assessed in the context of the tutorial: independent learning, group participation and problem solving. Again correlations were moderate but they found the highest correlation between peer and faculty ratings: independent learning (r=0.5); group participation (r=0.54) and problem solving (r=0.24) (all significant at p=0.01). In comparison the lowest correlation was between self and faculty ratings: independent learning (r=0.24); group participation (r=0.18) and problem solving (r=0.11) (all significant at p=0.05).

Bryan *et al*. (2005) found that students received significantly more positive comments from their peers than from themselves. Students were also ranked higher by their peers than by themselves with a mean (± SD (Standard Deviation) of 4.3 (± 0.5) and 3.6 (± 0.8) respectively, $p<0.001$.

Rudy *et al*. (2001) also present a number of possible explanations why students are more proficient in evaluating their peers in comparison to their own skills, knowledge and performance. Firstly students may be socially uncomfortable in presenting a wholly favourable impression of themselves to others and prefer to be modest in their self-assessments. Alternatively students at a certain level of training may have unrealistic goals and expectations of their abilities due to inexperience. Another possible explanation is a tradition of judgemental and punitive evaluation in medical education which inhibits students from expressing themselves. The way individuals judge performances may also go some way to explaining this anomaly in that they assess their peers at face value but apply global perceptions of performance to their own abilities. Finally the method of self-assessment may influence the outcome. For example a study which uses video recording may contribute to inaccurate self-assessment by causing anxiety and self-consciousness.

The general consensus here (albeit limited to three studies) that individuals are more able to accurately assess their peer's performance in comparison to their own is valuable when considering methods of validating self-assessment. The triangulation of a self-assessment measure by a more accurate measure should increase the value and meaningfulness of the exercise for an individual.

2.6.2 Individual Characteristics

A common aim of many studies was to identify factors and characteristics in individuals which would account for their differential ability to self-assess. There are two recurring themes which dominate the literature reviewed, namely gender and insight. There have been limited attempts to investigate the effects of cultural differences. Insight has become a field of study in itself as exemplified by the previously discussed work of Kruger & Dunning (1999). There is a separate section later specifically addressing insight. With reference to Kruger & Dunning (1999) insight may be defined as the ability to assess how well one is performing, when one is likely to be accurate in judgment and when one is likely to be in error. Experience is also considered later under the heading 'Factors influencing self-assessment'. Gender and cultural differences in self-assessment are discussed below from papers included in the review.

2.6.3 Gender

Researchers consider gender an obvious starting point in looking for potential reasons for differences in outcomes when individuals self-assess. There are more papers reporting differences in gender than any other type of sub-analysis. Despite this, the evidence drawn from across a number of studies is either inconclusive or contradictory.

Edwards *et al*. (2003) intentionally set out to investigate the influence of demographic factors on the accuracy of self-assessment. Given its clear objective to assess the influence of gender differences, and the sample size of the study (1,152 students over a 10 year period) the results of this study deserve credence. It was found in the study population of third year medical students in an obstetrics and gynaecology clerkship that men were 1.7 times (odds ratio (OR) 1.72: 95% confidence interval (CI) 1.53 to 1.93) more likely than women to overestimate their grades.

A similar conclusion was reached by Minter *et al*. (2005) who examined gender differences in surgical residents. The sample size was small (female n=10, male n=19)

but nevertheless the authors found that both male and female residents underestimated their abilities compared with faculty. In comparison female residents underestimated their abilities to a greater extent (-1.15 ± 0.42 points) than their male counterparts (-0.75 ± 0.19 points) but the difference between the two groups was not significant.

Bryan *et al*. (2005) in a study of 213 medical students found that males rated themselves more highly than females (mean 3.7 ± 0.8 (SD)) and 3.5 (± 0.9) respectively ($p=0.04$). Males received significantly more positive comments than females on peer evaluations (9.1 ± 2.5) and (8.4 ± 2.0) respectively ($p=0.025$) and were rated higher than females on peer provided numerical rating (4.4 ± 0.5 and 4.2 ± 0.5 respectively) ($p=0.02$).

In contrast, Leopold *et al*. (2005) discovered contradictory evidence on gender differences in confidence levels depending on when the measure was taken. They examined the confidence and self-assessment of 93 practitioners in performing a simulated knee joint injection. Measures of confidence and self-assessment were taken before and after they were randomized to three types of instruction: printed manual; video; hands-on instruction. The self-assessment was compared with objective performance standards measured by a custom designed knee model with electronic sensors that detected correct needle placement. Prior to instruction male participants were significantly more confident (6.32 points on a 10 point Likert scale) than female participants (2.95 points, $p<0.01$). In terms of performance there was no significant difference between the performances of men and women (6.62 and 5.86 points respectively, $p>0.05$). After instruction female participants were significantly more confident than male participants (8.77 compared to 6.98 points, $p<0.01$) and also had higher objective scores for performance (8.88 compared with 7.73 points, $p<0.05$).

Zonia & Stommel (2000) compared 73 interns' self-assessments of their medical knowledge and skills against those of their faculty, and stated that gender had no influence on either the interns' or faculty's ratings. However no data are presented in this brief research report to substantiate this conclusion.

Herbert *et al*. (1990) clearly set out to analyse the effect of gender on 142 third year

obstetrics and gynaecology students' assessments of their performance against grades assigned by different groups (faculty, residents) and using different methods (clinical activities, written exams, oral examinations). The authors concluded that in terms of both departmental ratings and self-ratings for all methods of evaluation there were no differences attributable to gender (range of p values 0.07 to -0.85).

Woolliscroft *et al*. (1993) attempted to identify the factors that influence third year medical students' (n=137) initial self-assessment of their clinical performance. Weak and negative correlations were found between self-assessments and college exam results but no statistically significant difference was found relating to gender (no p values presented).

Parker *et al*. (2004) looked at the ability of 311 family medicine residents to predict (i.e. self-assess) their performance on the in-training examination (ITE), regarded as an objective measure of medical knowledge. They found that residents demonstrated little ability to predict their examination scores (all Pearson correlations in 9 subject areas were less than 0.3) and there was no difference by gender.

Sommers *et al*. (2001) investigated how several variables including gender would affect physician faculty members' perceived self-efficacy for performing nine key professional role functions. They found that women (n=21) had lower self-efficacy scores than men (n=31) but that the difference was not statistically significant (p values ranged from 0.04 to 0.84 in the nine areas).

An example of contradictory evidence is found in the study by Evans *et al*. (2005). They examined the self-assessment skills of 50 surgeons in assessing their performance in removing a tooth. In using a checklist scale they found a significant difference between the mean scores of the assessors and male and female scores. Both males and females over-scored themselves compared to their assessors with males significantly more likely to do so than their female counterparts (difference in means (males – females) 1.94 (95% CI 0.26-3.62, p=0.03). However the same comparison with a global rating scale found no difference between males and females (difference in means (males – females) 0.09 (95% CI 3.36-3.55, p=0.96). In investigating reasons why individuals cannot assess they found no statistical difference between males and females on

either of the theories they were investigating i.e. impression management (trying to convey a favourable impression) and self-deception (lack of insight). However the authors recognise that the sample sizes were too small to provide definitive evidence (32 males, 18 females).

The number of studies analysing gender differences highlights the interest in this particular aspect of self-assessment. A number of studies found no difference in the ability of males and females to self-assess. However in terms of confidence there does appear to be a trend for males to express higher levels than their female counterparts. As with most research in this area however Leopold *et al*. (2005) found differing evidence depending on when the confidence measurement was taken. This study typifies the inconclusive nature of evidence in the analysis of gender differences which will no doubt continue to be a fertile ground for future research.

2.6.4 Culture

In comparison to investigations about the effects of gender (discussed here) and experience (discussed later under Clinical Skills), research into race and cultural differences is relatively scarce. Woolliscroft *et al*. (1993) correlated self-assessments and college exam results in third year medical students but found no statistically significant difference relating to race (no p values presented). Fitzgerald *et al*. (2003) concur that self-assessment accuracy is not related to ethnicity from a series of studies they have undertaken.

It is worth noting that the NHS ePortfolio could not be used to examine either gender or culture, as it could not contain this information about its users.

2.6.5 Insight

As outlined in the previous research section, a series of studies on psychology students (Kruger & Dunning 1999) explored the hypothesis that incompetent students over-estimate their ability because their incompetence denies them the ability to recognize competence or lack of it, either in themselves or others. The most competent students

tended to underestimate their performance, but improved their accuracy of self-assessment after benchmarking, whereas the less competent students tended to be more inaccurate after viewing others' performances. Increasing the competence of these students in logical reasoning increased the accuracy of their self-assessments, apparently by improving their meta-cognitive skills. Various researchers, including Hodges *et al*. (2001), have tested these hypotheses in clinical self-assessment settings.

Several of the higher quality papers reviewed addressed the relationship of the accuracy of self-assessment with competence, academic ability or insight into their performance.

Bryan *et al*. (2005) in a study of 213 first year medical students on an anatomy course stated that students with higher grades underestimated their own performance, whilst those doing poorly tended to overestimate their performance. They did not provide figures to substantiate this assertion, but did find that self-rating scores were weakly positively correlated with the final grades ($r=0.14$, $p=0.04$).

Edwards *et al*. (2003) asked third year students on an obstetrics and gynaecology clerkship to estimate their final examination and clerkship grades at the beginning of the clerkship, and again just prior to the final examination. Complete sets of grades and predictions were obtained from 1139 students out of 1152. Students were more likely to accurately predict their clerkship grade than their examination grade, but for both estimates, the students ranked in the lowest third were more likely to overestimate their grades than those in the top third, who tended to underestimate their grades. The logistic regression results with 'overestimate' as the modelled outcome give odds ratios of 4.38 (CI 3.79–5.06) for lower versus upper third of students, and 1.90 (CI 1.66–2.18) for middle versus upper third of students.

Parker *et al*. (2004), asked 311 family medicine residents to estimate their performance in nine content areas of an in-training examination. They also found that high scorers tended to underestimate their scores and low scorers to overestimate them. The most accurate predictions were made by the students in the middle two quartiles.

Leopold *et al*. (2005) examined the confidence and self-assessment of performance of

93 practitioners attending an educational session on knee injection, in relation to assessment by trained observers. Their rationale was that professionals must decide whether they have the competence to undertake a procedure, and that this decision is based on their level of confidence, as well as their background, education and skill. They found an initial significant but inverse relationship between confidence and an objective measure of performance before instruction ($r=\pm0.253$, $p=0.02$), that is greater confidence was associated with poorer performance. They also found that confidence before instruction was strongly and directly correlated with the participants' assessment of their own performance ($r=0.42$, $p=0.001$ and therefore concluded that confidence was associated with overestimation of self-assessed performance. The effect of instruction on self-assessment was also measured and this is described in the relevant section below.

In a study of 25 resident physicians (Millis *et al*. 2002) self-assessment scores for an interview with a standardised patient (SP) were compared with those of the standardised patients and those of faculty. There was reasonable correlation between faculty and standardised patient ratings, (0.50, 95% CI 0.16-0.73) but lack of correlation between standardised patient and physician self-ratings (0.11, 95% CI -0.28-0.47). The resident physicians who were rated poorly by the SPs tended to rate themselves as high as physicians who were highly regarded by the SPs.

Woolliscroft *et al*. (1993) examined the clinical self-assessments of 137 out of 142 third year medical students compared with external measures of performance including the Medical College Admission Test (MCAT) and students' college grade-point averages (GPAs). Students in the lowest quartiles for both the GPAs and MCAT scores rated themselves highest for all skills except application of knowledge, for which students in the top quartile had a higher mean.

Mandel *et al*. (2005) compared the self-assessments of 74 out of 92 surgical residents with faculty ratings on two assessment measures, open surgical skills and an external global skills checklist. There was a high correlation between residents and faculty ratings on specific tasks and global skills. Unlike other studies in this section, these authors did not find that residents with poor skills were unaware of their deficiencies.

The literature reviewed contains several instances of over-estimation by poor performers, and under-estimation by those who perform well. These studies reinforce the ideas of Kruger and Dunning who argued that those who lack competence also lack the meta-cognitive skills (i.e. the awareness or ability to analyse one's own thinking and learning processes) to recognise their poor performance. Dunning (2006) explores this idea in more depth in a recent paper, suggesting that ''people misjudge their incompetence not because of a lack of honesty with themselves, but rather because of a lack of the essential cognitive tools needed to provide correct self-judgments''. An alternative explanation might be that such results merely reflect poor correlations between self-ratings and faculty or other assessments. Hence, rather than drawing on a psychological defence mechanism to account for the discrepancy between different raters, this finding could indicate a central tendency or regression to the mean in self-assessments. It is interesting, however, that in the Mandel *et al*. (2005) study it was in the area of practical skills in which the poorer performers' estimates correlated with faculty ratings and with higher scorers' estimates. This will be discussed further in the section on practical versus cognitive skills.

2.7 External Factors

A variety of factors outwith self-assessment was examined to determine whether they had an impact on the process.

2.7.1 The Purpose of the Self-Assessment Task

In In reading of the literature it became clear that authors seldom gave information on whether or not participant self-assessment contributed to the final marks of the student or if the student self-assessment was seen by the tutor/external assessors prior to their mark being attributed.

This is important as in the first of these scenarios there may be pressure on the student to inflate their marks in order to improve their grades, reducing the apparent accuracy of their self-assessments. The impact of the second is more complex, some

may see their self-assessment as a means of pressuring their tutor into giving a higher mark (it may be easier for a tutor to give a D to a student who self-assesses as D rather than one who self-assesses as B) while others may be too modest to suggest a high score even if they think they might achieve it.

Only one high quality study was found exploring the impact of either of these arrangements. Evans *et al*. (2005) explored the possible influence of self-deception as a possible reason for the discrepancy between self (surgeons') and assessors' ratings. They asked dental surgeons to rate their skill following removal of a third molar observed and rated by two assessors (who had good inter-rater reliability) and in addition the Paulhus Deception Scale 7 (PDS) (Paulhus 1998) was simultaneously administered. This is a validated 40 item questionnaire that measures an individual's tendency to give socially desirable responses on questionnaires. There are two components of this scale, Impression Management (IM) and Self-Deception Enhancement (SDE). Impression management refers to the tendency to give inflated self-descriptions by 'faking or lying' and to deliberately convey a favourable impression ('faking good') whereas self-deception enhancement indicates overconfidence and lack of insight. Seventy per cent of surgeons had impression management scores suggesting that they may have been deliberately trying to give a favourable impression. These IM scores correlated significantly (r=0.45, p=001) with the inability to assess their own surgical skills. Although 30% of the surgeons in this study showed lack of insight, that is to say they scored high or very high for self-deception enhancement, there was no evidence to suggest this affected their opinion of their surgical performance. It could be speculated that this could be influenced by professional culture and/or conditions specific to the research environment.

Further research exploring the impact of the purpose of self-assessment on its accuracy is required. Additionally research is needed to explore the impact of student self-assessment on external assessment.

2.7.2 Practical Skills Versus Theoretical Knowledge

Few studies have specifically set out to determine if self-assessment of cognitive skills

differs from that of practical skills.

Edwards *et al*. (2003) compared the self-assessment skills of obstetrics students and found that a higher proportion of students were able to predict their clerkship grades (based on performance) than their grade by examination (56% v 31% at the start of the attachment and 61% v 32% at the end, both $p<0.001$). However, Fitzgerald *et al*. (2000) compared self-assessment of two sets of skills, which they described as cognitive (chest-pain questions, EKG analysis, x-ray analysis) and performance (examination of breast, chest pain patient, unconscious patient, paediatric examination, communication skills). They found no difference in accuracy of self-assessment between either type of task.

Additionally there is evidence from other good quality studies which seems to show that practical tasks, particularly surgical tasks, appear to be amenable to self-assessment especially if feedback on performance is included. The review found several papers which suggested that students had at least moderate skill in self-assessment of performance or practical skill.

Woods *et al*. (2004) surveyed 266 American physicians about their ‘‘comfort’’ (assessed on a 4 point scale) with differentiating between smallpox and chicken pox and tested them with a simple 4 question knowledge test and a visual diagnosis using photographs. 178 physicians responded. In logistic regression controlling for predictive variables (general experience, experience of rashes and speciality) only ‘comfort’ in diagnosis was predictive of knowledge of small pox diagnosis (OR 2.2, 95% CI 1.4–3.3). No parameter was found to be predictive of performance in identifying smallpox from photographs.

Ericson *et al*. (1997) found that dental students using performance guidelines in the area of cariology (1,373 diagnostic, preventative and restorative procedures) agreed with their tutors in 87% of assessments.

Ward *et al*. (2003) in a small study explored the self-assessment skills of 28 senior resident surgeons in laparoscopy. They demonstrated a correlation of $r=0.50$, $p<0.01$ immediately after conducting the surgical procedures which rose to $r=0.63$, $p<0.01$ after review of their videoed performance.

Similarly Mandel *et al*. (2005) compared self-assessment of proficiency on a variety of surgical bench procedures with the reliability-tested Objective Structured Assessment of Technical Skills (OSATS) in 74 obstetrics and gynaecology residents. They demonstrated high correlations with both open procedure skill ($r=0.74$, $p<0.001$) and laparoscopic skills ($r=0.67$, $p<0.001$).

Evans *et al*. (2005) showed modest agreement (intra-class correlation co-efficient of 0.51) between assessors and fifty dental surgeons completing a checklist on performance of extraction of a mandibular third molar.

Lane & Gottlieb (2004) compared fifty third year medical student self-assessments of interviewing skills using a 21-item five point self-assessment scale with two faculty members' assessments. Medical students disagreed with faculty in their assessment 14% of the time, but this reduced to 7% following feedback.

Weiss *et al*. (2005) examined the self-assessment skills of 47 third year medical students on an obstetrics and gynaecology rotation. Skills were examined in five areas: fund of knowledge, personal attitudes, clinical problem solving skills, written/verbal skills and technical skills. Self-assessments were correlated with exam results and faculty and resident ratings. They found a statistically significant weak to moderate, positive correlation between students' self-assessment and final clerkship grade for written/verbal skills ($r=0.390$, $p=0.002$). A statistically significant agreement between raters was also revealed for written/verbal skills ($p=0.003$). Weak, non-statistically significant, positive relationships were revealed for fund of knowledge, clinical problem-solving and technical skills. A weak, negative, non-significant relationship was revealed for personal attitudes, and there was no statistically significant relationship between students' prediction of their exam score and categorized true score ($r=0.49$, $p=0.717$). This leads the authors to conclude that at the end of their obstetrics and gynaecology clerkship, third-year medical students are better at assessing their technical and written/verbal skills than their global fund of knowledge and personal attitudes.

Leopold *et al*. (2005) explored the impact of education and feedback on self-assessment of skill in the performance of a simulated knee joint injection. Ninety three

practitioners were randomised to receive skills instruction through a manual, a video or hands-on instruction. Each participant performed one injection before and after instruction. All participants completed pre and post-instruction questionnaires on confidence and provided self-assessments of performances before and after instruction. Before instruction, participants' confidence was significantly inversely related to competent performance ($r=-0.253$, $p=0.02$). After instruction, performance improved significantly in all three training groups ($p<0.001$) with no significant differences in efficacy detected. After instruction, confidence correlated with objective competence in all groups ($r=0.24$, $p=0.04$); however, this correlation was weaker than the correlation between the participants' confidence and their self-assessment of performance ($r=0.72$, $p=0.001$).

In contrast to this, however, Rudy *et al*. (2001) showed poor correlation ($r=0.19$, NS) between self and faculty assessment in communication and interviewing skills in 97 first year medical students (although good correlation $r=0.50$, $p<0.0001$) between faculty and peer assessment of the students).

Antonelli (1997) showed relatively good correlation ($r=0.49$, $p=0.0006$) between global self-assessment of skill in second year medical students and perceptors' final grades but confidence in self-assessment skill was not correlated with accuracy of self-assessment. Students in this group, however already had received two thirds of their year examination results and so were in a good position to predict their final score.

However, there were five included papers that failed to find a correlation between self and external assessment of knowledge in the areas of:

- medical knowledge (self-assessment versus the In-training examination) in residents in family medicine (Parker *et al*. 2004),
- assessment of performance in undergraduate PBL tutorials (Sullivan *et al*. 1999; Reiter *et al*. 2002), general practitioner knowledge of thyroid disorders and diabetes (Tracey *et al*. 1997),
- general practitioner knowledge of techniques for assessing evidence based medicine (Young *et al*. 2002),
- residents' knowledge of critical care as assessed by MCQ (Johnson & Cujec

1998).

Fitzgerald *et al*. (2003) report a longitudinal study of medical students' self-assessment ability over three years. They noted this deteriorated in the third year. However, the examination format, which was OSCE based, was considerably different from traditional knowledge based exams they had previously sat and the authors posited that rather than the deterioration in self-assessment ability being due to increasing experience, it was due to the format of the examination.

It is not clear why practical skills may be better self-assessed than knowledge, but it could be that their outcomes are harder to dispute so the potential for self-deception about one's abilities is less. For example, it is harder not to recognise when a clinical procedure has gone poorly, especially when immediate feedback might be forthcoming from colleagues and the patient. This may not apply, however, to interpersonal skills which seem relatively poorly self-assessed in the absence of structured feedback, as the individual can more readily deceive themselves as to the outcome.

2.8 Factors Influencing Self-Assessment

The review found a number of factors that could affect self-assessment which are listed below.

2.8.1 What Factors Can Improve the Development of Self-Assessment Skills?

This section focuses on studies which report that self-assessment skills can be improved. Kruger and Dunning (1999), already referred to above, involved a series of psychological experiments in which they identified that people vary in their ability to self-assess. Of particular importance are the two groups who either over-rate or underrate themselves. Those in the top quartile who under-rated their abilities were able to improve their self-assessment rating when shown the results of other people's work. This process helps the able student to benchmark their ability in relation to the

ability of their peers, resulting in a more accurate self-assessment. The improvement in the accuracy of self-assessment has only been demonstrated for able students who previously under-rated their performance. Kruger and Dunning noted that students in the bottom quartile consistently overrated themselves despite any benchmark feedback. Self-assessment in this group was improved only by educational input to increase the level of knowledge. Thus level of knowledge or skills needed to be raised in order to improve the accuracy of self-assessment.

2.8.2 Video Feedback and Benchmarking

The importance of feedback as a tool to increase the accuracy of self-assessment was referred to by Gordon (1991). Ward *et al*. (2003) reported on whether self-assessment accuracy improved following video feedback after completing a surgical procedure and comparing it with a validated gold standard of expert raters. The 26 surgical residents rated their performance immediately after completing the surgical procedure. Their ratings were moderately correlated with the expert ratings ($r=0.50$, $p<0.01$). The correlation increased significantly after the residents viewed a video of their performance and then repeated the self-assessment ($r=0.63$, $\Delta r=0.13$, $p=0.01$). This study does suggest that viewing one's own performance and then completing a self-assessment is more accurate than merely relying on recall of one's own performance. Then the authors asked the residents to view four videos that represented a range of abilities, thus providing benchmarks for each level of skill. The authors expected that knowing what the standard looked like at each level would lead to a further improvement in the self-assessment accuracy of the residents' own level of skill. However no further improvement was identified and the authors postulated that this may be due to the senior skill level of the surgical residents who would have already had a good knowledge of the range of levels of performance. The margin for further improvement therefore in these circumstances would have been too small to detect a significant difference.

A similar study using benchmarks was conducted by Martin *et al*. (1998). The study involved 25 first and 25 second year family residents. The residents were observed by

two experts while conducting a complex consultation with a standardized patient about suspected child abuse. The experts assessed the residents and the residents self-assessed their performance using the same scale. The residents were then asked to assess four benchmarked performances to determine whether the residents could identify the different benchmarked performances and whether they would match expert opinions. Following the benchmark tasks the residents were asked to reassess their own performance. The first self-assessment had a low correlation with the expert rating ($r=0.38$), but the correlation with experts increased significantly ($p<0.05$) after viewing the videos and re-assessing themselves ($r=0.52$). The change in self-assessment after viewing benchmarked performances brought the assessments closer to the ratings used by experts, suggesting they were using the scale in a similar way. The mean resident–expert correlation on the benchmarked tapes was quite high (0.72) but there was quite a wide range (0.57 to 0.89). Further analyses found that the ability to correctly benchmark the videos was not related to either the ability to perform the task or the ability to accurately self-assess.

2.8.3 Video and Verbal Feedback

Lane and Gottlieb (2004) videoed the performance of 60 students conducting medical interviews and then asked students to self-rate their performance on a Likert scale that covered 21 key elements. The authors reported that the trend was for performance to improve from first to second time (319 of 432 instances, or 74% of the time). Also agreement between the rating of the tutor and those of the students improved on the second performance (14% down to 7% of errors) with a significant decrease in the rate of inaccurate assessments ($p=0.001$). Feedback from the tutor and from viewing oneself perform was identified as the stimulus for the improvement in performance. The increase in agreement on the rating scale was again linked to feedback from the tutors who gave their views on how good the performance was and why, thus enabling the student to recalibrate what a good performance would look like. This falls in line with other findings that demonstrable skills are better self-assessed, particularly with structured feedback. Given the ease and prevalence of video technology, once ethical

considerations are taken into account, the use of video could become a powerful tool used in conjunction with self-assessment and is an area that will certainly be attractive for future research.

2.8.4 Instruction

Leopold *et al*. (2005) conducted a before and after study with 93 practitioners who were randomly assigned to receive one of three instructions to improve skills on giving a knee injection. The three types of instruction were: printed manual, video and hands-on instruction. The practitioners completed a self-assessment before and after the interview. Before the intervention increased confidence was related to poorer performance ($r=-0.253$, $p=0.02$). After the instruction performance improved significantly in all groups ($p<0.001$), but there were no significant differences between groups. The correlation changed after the intervention from a negative to a positive correlation, showing that confidence was related to performance, but the correlation was weaker ($r=0.24$, $p=0.04$). The authors concluded that even low intensity forms of instruction improved confidence, competence and self-assessment.

2.8.5 Experience

There is some evidence that increased experience in a skill or knowledge is also reflected in higher scores on a self-assessment scale. Studies examined two particular aspects of experience. The first is the relative level of experience of the participants in relation to their clinical knowledge, skills or expertise, for example novice versus expert. Typically this might involve first year undergraduates being compared to third year undergraduates. The second aspect of experience explored is the effect of exposure on an individual's ability to self-assess. This involves examining proficiency before and after an intervention or experience e.g. attendance on a rotation. The objective is to determine whether exposure to a skill or experience increases an individual's accuracy in assessing their performance as they become better accustomed to the respective task or skill and acquire better knowledge.

2.8.6 Novice Versus Expert

Wilkerson *et al*. (2002) investigated the effects of an enhanced curriculum in cancer prevention on medical students' (n=333) knowledge and self-perceived competency in the use of counselling and screening examinations during the first three years of medical school. This enabled them to compare the three different years of students with varying levels of knowledge and experience. They reported that students' knowledge of cancer prevention significantly improved over time (e.g. third year students scored significantly higher than the years below them, $p<0.001$). The reported improvement in the self-assessed skills of counselling and screening skills was correlated to hands-on practice. When practice was removed, as in the second year, the improvement in self-assessed skills was absent. This finding suggests that hands-on practice provided an opportunity for knowledge and skills to be tested out and providing the individual with some feedback increased the self-rated competencies.

Herbert *et al*. (1990) evaluated the effect of previous clerkship experience on the actual grades that 142 third year students achieved on a six week obstetrics and gynaecology clerkship. There was no correlation between the grades achieved and previous clerkship experience and more experience did not affect students' ability to self-assess. Unfortunately no data is presented to verify this conclusion.

Sommers *et al*. (2001) specifically examined the length of faculty members' (n=54) experience on their self-perceived efficacy for carrying out key medical functions. They concluded that time in faculty did not have any significant effect on the total self-efficacy scores for the nine professional role functions examined i.e. increasing the length of time in a faculty position did not influence self-efficacy scores (p values ranged from 0.042 to 0.78 in the nine areas). Furthermore they found no statistically significant association between age and the total self-efficacy score or that for the nine individual areas investigated (no data are presented to verify this finding).

Leopold *et al*. (2005), also reported that prior to the intervention, practitioners with more expertise rated themselves higher than their peers, although their performance was not significantly better. After the intervention there was again no correlation with experience and greater performance (as measured by increased years in practice or by

giving three or more injections).

Paradise *et al*. (1997) asked 206 physicians who rated their skills as above average in evaluating cases of suspected sexual abuse to examine seven simulated cases by means of a questionnaire. The physicians' descriptions and interpretations of the simulations were compared with consensus standards developed by an expert panel. In three of the simulations the most experienced physicians resembled the panel more closely than did the less experienced ($p \leq 0.001$). This leads to the conclusion that among physicians who self-rate themselves as skilled, assessments made by more experienced physicians may relate more closely to consensus standards than those made by less experienced physicians.

2.8.7 Exposure and Feedback

Edwards *et al*. (2003) conducted a before and after study involving 1,152 students comparing the differences between predicted and actual final examination and clerkship grades. This was an extensive study over ten years of third year students (n=1,152) in an obstetrics and gynaecology clerkship. Students were more likely to correctly predict their clerkship grade than their examination result, at the beginning (56% vs 31%, $p<0.001$) and at the end (61% vs 32%, $p<0.001$). The authors reported that students who had slightly shortened placements (6 weeks compared with 8) were 3.6 times more likely to overestimate their clerkship performance than the students on the 8 week placement. Also students who did the clerkship earlier on in their careers (during the autumn semester) were 1.55 times more likely to overestimate their performance than those who did it later on in the spring semester. The authors suggest that on-going feedback during the clerkship may have had an effect on the greater predicted accuracy of the clerkship grade compared to the exam grades. The authors postulate the importance of feedback, which they suggest plays a mediating role in accurate self-assessment.

Zonia and Stommel (2000) evaluated the difference between interns' self-assessments (n=73) and those made by their faculty. In terms of experience they found that interns' self-ratings and equivalent faculty ratings consistently increased in the first five

months of their rotations (p=0.001). However after the fifth month the ratings reached a plateau.

Gruppen *et al*. (2000) ran a study which aimed to correlate how amounts of study time linked to changes in self-assessed diagnostic capabilities over the course of a three month clerkship. The subjects were 107 medical students in three consecutive cohorts of an internal medicine clerkship. This was a before and after study which correlated a self-assessed measure of confidence at the start and finish of the clerkship with an estimate of time spent studying respective topics. The researchers found a modest but positive correlation (mean co-efficient=0.25, SD=0.20; 95% CI 0.21 to 0.29) leading them to conclude that spending more time on a given topic resulted in an increase in self-assessed diagnostic skill for that subject. They cautioned that individual variation influenced the strength of the relationship, it being much stronger for some students than others (range=-0.23 to 0.89).

Eva *et al*. (2004) in a study of 265 Canadian medical students found no evidence that performance in self-assessment improved over 2.5 years of schooling. They did find that students who estimated their examination performance after sitting the examination were more accurate than those who predicted their score before taking the examination.

The level of experience of those self-assessing raises an interesting question in the literature, namely whether it is experience in the knowledge or skill being assessed that determines self-assessment ability or experience of self-assessment itself which is most important in determining accuracy. Ward *et al*. (2003) examined the self-assessment accuracy of 26 surgical residents and whether self-observation of their performance by video and the opportunity to view benchmark videos of performance would improve their self-assessment ability. Initially there was a moderate correlation between experts' evaluations and residents' self-evaluations ($r=0.50$, $p<0.01$). They found that self-observation did improve self-assessment ability ($r=0.63$, $\Delta r=0.13$, $p<0.01$) but exposure to benchmarked performances did not ($r=0.66$, $\Delta r=0.03$, NS). This leads them to conclude that ability to self-assess is related in this case to surgical experience rather than self-assessment experience.

In summary, these studies highlight the importance of both feedback on performance, and of increasing knowledge of the task to increase understanding and recalibration of what a good performance involves.

2.8.8 Perceptions and Attitudes Towards Self-Assessment

The review set out to determine the attitudes towards and perceptions of learners and teachers to self-assessment. However, few papers in the review made more than a passing reference to this feature of self-assessment and, among those that did, no single paper met the quality threshold for inclusion. There were no studies that focused on perceptions alone; these were always of secondary consideration.

Whilst the evidence is not robust, the papers examined would seem to suggest a favourable response towards self-assessment activities on the whole by participants. There is occasional indication of stressful reactions experienced by students in some studies but this requires further exploration.

The acceptability of self-assessment as an educational tool is assumed rather than explored in the literature. There is an urgent need for high quality research in this area. The lack of a robust evidence-base about attitudes towards self-assessed activities is somewhat contrary to their importance in practise for identifying leaning needs and maintaining competence in health professional behaviour. The dearth of robust qualitative research is of particular concern in this field.

2.9 Discussion

The research questions addressed by this review sought evidence for the effectiveness of self-assessment interventions to:

- improve the accuracy of learner perception of their learning needs,
- promote an appropriate change in learner learning activity,
- improve clinical practice,
- improve patient outcomes.

Subsidiary research questions addressed factors affecting the accuracy of self-

assessment, and learners' and teachers' perceptions of and attitudes towards self-assessment.

Overall, it appears that the review, despite a robust methodology, was largely unable to answer the specific research questions, and provide a solid evidence base for effective self-assessment. No papers were found which satisfied Kirkpatrick's hierarchy above level 2, and found no studies which looked at the association between self-assessment and resulting changes in either clinical practice or patient outcomes.

However, in terms of the subsidiary questions, while no indisputable evidence was found, the systematic review did identify several factors which appear to influence self-assessment. In order to increase the understanding of the conditions which are associated with accurate self-assessment, it is recommended that these areas would merit further research.

2.9.1 Findings

An important conclusion across a number of studies was that individuals are far more able to accurately assess their peers' ability than their own. Peer assessments also appear to be more in line with faculty assessments of performance than self-assessments. This could be important when considering methods of validating self-assessment.

Ability and experience would appear to have some impact on self-assessment, with several papers exploring the relationship between accuracy of self-assessment and competence or academic ability. The findings from these studies broadly support the idea that competent practitioners are reasonably accurate in their self-assessment, and it may be possible to improve this accuracy. On the other hand, people who lack competence are less likely to be aware of their deficiencies as evidenced by self-assessment, and to be less responsive to strategies for improving accuracy. This has important implications, critically for under-performing health professionals, and is worthy of further research.

There is some evidence from the review that practical skills may be better self-assessed than knowledge. As noted in the results section, this could perhaps be

explained by the fact that the outcomes of practical skills are harder to dispute and so the potential for self-deception about one's own abilities is less. Observable performance also lends the opportunity for direct feedback.

The importance of feedback and benchmarking has been identified in a small number of studies in the review as increasing the accuracy of self-assessment by increasing the learner's awareness of the standard to be achieved.

Many studies used gender as a starting point in looking for potential reasons for differences in self-assessment outcomes. Although there were more papers examining differences by gender than any other type of sub-analyses, most of the evidence here was inconclusive or contradictory and may have been relative to the type of activity under consideration.

There was no high quality evidence to suggest that race or culture impact on an individual's ability to rate themselves objectively.

In the context of how self-assessment is perceived by learners and teachers, the review suggests that the acceptability of self-assessment is seldom explored. Of those which did address this, there would seem to be a favourable response to self-assessment activities by participants, although self-assessment may be stressful for some students and even potentially threatening. Attitudes towards self-assessment may be influenced by the purpose of the self-assessment activity, that is whether self-assessment is undertaken for formative or summative outcomes. The need for high quality research is particularly urgent in this field.

2.9.2 Strengths of the Review

At the start of this research, considerable time was spent developing a rigorous methodology with which to conduct the review. Agreeing an explicit definition of self-assessment was itself a complex activity and this will be addressed later.

- As noted in the Methods section, a rigorous review process was developed, which incorporated several iterative stages.
- Development and use of a standardised coding and quality
- Checklist adapted from validated tools

- All papers were reviewed independently in duplicate, with recourse to an adjudicator to resolve disagreements
- Iterative process of reviewing and discussing papers and if necessary revisiting the full text
- Regular discussion between pairs and with the whole group to clarify concepts
- Peer review/feedback from presentations at international conferences (ASME, AMEE and Ottawa conferences).

2.9.3 Hindrances

Some 'teething problems' were experienced, perhaps inevitably, around the development phase of the electronic coding form. Overcoming these has benefited a subsequent review which is using a similar e-form.

Although a large number of papers resulted from this original search (n=5,798), only a small proportion were of sufficient academic rigour to be included in the review (n=32). Research on self-assessment has been fraught with methodological problems, and this is reinforced by the review where reasons for exclusion included no clear definition of self-assessment, inadequate information on sampling strategies, and insufficient reporting of methods and analysis. Similar concerns about the quality of published research in self-assessment have been expressed by Davis *et al*. (2006). These authors conducted a more focused review, limited to a comparison of physician self-assessment with observed measures of competence. Despite this more specific context, only 17 out of 725 papers met all the inclusion criteria. One of the implications from both reviews is that the peer review process in many journals may need to be more rigorously implemented.

Most of the papers of sufficient quality to be included in the review concentrated on judging the accuracy of self-assessment by comparison with some external standard (as was the focus of the Davis review), but as outlined above there are problems with this approach. This left few papers selected for the review that actually addressed the specific research questions.

Self-assessment, no matter how it is defined, is a complex concept which does not lend

itself to objective measurement. It may be, therefore, that the conclusions were limited by the definition of self-assessment, and that the outcome of the review would have been more definitive if it had used a broader definition, particularly one which takes account of meta-cognitive skills. Despite attempts to standardise the approach to inclusion and exclusion of papers, there is inevitably a subjective element to making this final judgement, and this may have resulted in some borderline papers being excluded. The risk of this would have been mitigated by an agreement by all reviewers to include papers that were judged to be on the cusp of inclusion.

2.9.4 Philosophy of Self-Assessment and Problems of Definition

Self-assessment was defined as "a personal evaluation of one's professional attributes and abilities against perceived norms".

Very few of the papers that were reviewed defined the concept of self-assessment that they were researching. The majority of them set out to determine the 'accuracy' of self-assessment in terms of quantitative comparisons with external measures or 'expert' ratings. Ward *et al*. (2002) point out the problems with these types of studies, namely lack of validity and reliability of the 'gold standard', the likelihood of differential use of scales among students, and problems of group level analyses.

Colliver *et al*. (2005) concur with Ward *et al*. (2002), and go further in suggesting that this type of quantitative analysis of 'guess your grade' type studies is not relevant to the daily ongoing self-assessment of practice. The latter involves the recognition of specific deficits in knowledge or skills in the context of the clinician's practice. They make the point that self-assessment for ongoing self-directed learning is a qualitative exercise, concerned with specific subjects in an individual context. This would lend itself to a narrative approach about an individual's clinical knowledge and skill, and indeed could not be quantified. They suggest that this personalised assessment in practice should be the target of research, and that this is beyond the conventional quantitative research paradigm of academic reflection in the published literature.

Eva and Regehr (2005) follow a similar thread when they argue that although simple definitions of self-assessment are attractive, they tend to cause difficulties because

they do not allow for the complexity of the concept. They suggest the adoption of a different paradigm, in which professionals constantly self-assess in terms of their own strengths and weaknesses in relation to situations that they experience. The ability to identify one's weaknesses can lead to knowing when to ask for help with a case, or to setting appropriate learning goals. Being aware of one's strengths allows one to persevere with a correct course of action despite initial setbacks, and to set realistic, challenging, but achievable learning goals.

The authors point out that self-concept, ''a relatively sweeping cognitive appraisal of oneself'', and self-efficacy, ''a context-specific assessment of competence to perform a specific task''' will both influence self-assessments. They argue that self-efficacy differs from self-assessment in that it influences our performance, a strong sense of self-efficacy leading to a greater chance of success.

In the introduction, some reference was made to how self-assessment was defined for the review, and the difficulty this raises in the context of self-referent thinking. Wooliscroft *et al*. (1993) draw on psychological literature to argue that an individual's view of self, or 'self-concept' results from external feedback and introspection. Accurate self-assessment clearly depends on congruence between self-representation and reality, but these authors argue that over time, self-representation becomes increasingly resistant to change despite feedback. This reinforces Gordon's (1991) finding that self-assessment did not always change as a result of external evaluative information. It is not clear however why low achievers are more likely than high achievers to overestimate their abilities, although some authors suggest some kind of psychological 'defence' mechanism (Woolliscroft *et al*. 1993). Such psychological self-protection strategies could also explain the studies that found that generally we assess others more accurately than we assess ourselves.

In the psychological literature, the concept of self-efficacy originates from a theoretical basis which emphasises the importance of feedback in shaping subsequent action (Bandura 1977, 1986). Like Woolliscroft's explanation of self-representation, self-efficacy thus incorporates environmental (external) and cognitive (internal) factors on learning behaviour. Eva and Regehr (2005) have defined self-efficacy as ''an

individual's judgement of her capabilities to complete a given goal'' (p. 548). These authors argue that the literature on self-assessment focuses on 'accuracy' (reinforced by the review) while research around self-efficacy focuses on the consequences of particular self-efficacy beliefs and their impact on future performance of tasks, which is arguably a key outcome. They also address the need to consider a third source of variation in self-assessment capacity, namely the meta-cognitive factors which affect individual judgements about learning, and in particular how individuals process the feedback and judgements about their performance made by others. As already noted, Kruger and Dunning (1999) hypothesised that deficient self-assessment may result from lack of meta-cognitive skills, and cited some evidence that improving meta-cognitive skills (in this case logical reasoning) improved self-assessment accuracy. Eva and Regehr (2005) have reviewed the research paradigms of several different but related disciplines. They express the view that the literature on reflective practice supports the idea of moving away from the concept of self-assessment as a 'conscious meta-cognitive and usually post-hoc summative process', and that safety in professional work requires that self-assessment be conceptualised as an ongoing 'reflection-in-action', constantly monitoring one's ability to deal with the emerging situation.

In a paper published since the review commened, Dunning (2006) argued that the flawed nature of self-assessment could result from individual cost/benefits analysis – a theory well-documented in the context of risk-taking health behaviours. Strategies suggested for correcting mistaken self-judgements include recognising the importance of listening to external feedback, especially from peers, or improving meta-cognitive skills to be more realistic in the light of external 'evidence'. The third strategy proposed by Dunning is simply to adopt 'cognitive repairs'– in other words recognise that self-assessment is often inaccurate, and make appropriate allowances.

The accuracy of self-assessment as a measure of clinical performance may in fact be no worse (and no better) than any other single judgement of competence. There is a large body of evidence to suggest that many judgements (and methods) are required before stable and reproducible ratings of performance can be obtained (Carline *et al*. 1989;

van der Vleuten & Swanson 1990; Williams *et al*. 2003). Perhaps the nature of the self-assessment task is the issue here. In setting appropriate goals for learning, individuals must be able to identify their own weaknesses as well as their own strengths in the context of good professional practice. Relying solely on a self-assessment tool may be insufficient to determine the full extent of learning needs. In a paper already referred to earlier in this review, Amery and Lapwood (2004) found a clear disparity between respondents' self-rated skills and their educational requirements as derived from personal diaries. The gap between perceived and actual need led these authors to make a case for multiple assessment tools to fully identify the ongoing training required by health professionals.

In this study, the use of self-assessment as a single measure failed to pick up unmet educational, training and support needs in areas of clinical practice. But to discount self-assessment as wholly inaccurate or flawed, however, is rather to miss the point. We should be aware of the limitations of self-assessment but use it alongside other sources of information to provide broader, more holistic assessments of competence and learning activity for health professionals in practice. An opportunity to do just this became available with the creation of the ePortfolio and its collation of training and assessment data.

2.10 Conclusions

Self-assessment is integral to lifelong learning in the health care professions. However there is evidence that in some contexts and tasks self-assessment is inaccurate. More worryingly there is evidence that those who are least able are also least able to self-assess accurately. If self-assessment is to remain the cornerstone of continuing professional development and in determining how regulatory appraisal requirements are to be met, we need to have a greater understanding of what forms of self-assessment are useful in determining learning needs, and what impact these have on future learning activities.

The systematic review has been unable to answer these questions, but it has added weight to the arguments to consider different research paradigms to significantly

increase the understanding of how self-assessment works or can be improved. The review did however find themes in the literature around self-assessment that offer clear possibilities for future research to increase the understanding of the process. Based on this work, it was decided to take the review's questions forward within a far more detailed and comprehensive set of data – the training data of an entire year of Foundation doctors that follows below in the case study of Chapter 4.

This review was published as "The effectiveness of self-assessment on the identification of learner needs, learner activity, and impact on clinical practice" (Colthart et. al, Medical Teacher, 30 (2) 2008).

2.11 Update Search

Because of the time elapsed since the last search of the evidence base (during the second systematic review), an update search was conducted on the self-assessment from 2006 until end of November 2008 to examine whether more recent papers answered the untested questions and/or confirmed existing findings. The search employed the same strategies across the same databases as were used in the full and updating search within the systematic review itself.

The search resulted in 704 unique hits. The titles and (where available) abstracts were scanned for relevance and 47 of these were retained for a close reading/retrieval of the abstracts. From the 47, twenty eight were judged to be potentially relevant and were retrieved in full text.

The 28 papers can be thematically linked to the systematic review's research questions: *Is self-assessment effective in improving perception of learning needs?* (5), *Is self-assessment effective in promoting change in learning activity?* (2), *Is self-assessment effective in improving clinical practice?* (13). The remaining eight papers did not directly address the research questions but were considered for potential

relevance in informing general discussion.

The reviewed papers from the update search confirmed evidence found in the systematic review, in that feedback (particularly video) was seen to improve self-assessment, self-assessment ability is not developed through a curriculum and clinical skills were more accurately assessed by self than “softer” skills. Four papers also set out to examine Kruger and Dunning’s findings within a healthcare context and each reported a confirmation of their seminal work. The update search did not reveal anything to challenge the conclusions, or gaps in the evidence base, of the initial systematic review.

2.12 Future Research: Self-Assessment

From the discussion above and the findings of the review, some of the review group felt there was a need for a move away from quantitative comparison studies of the ‘accuracy’ of self-assessment. As Eva and Regehr (2005) point out, the problem with this paradigm runs deeper than flawed methodology of studies. They suggest that the problem is one of ‘‘a failure to effectively conceptualise the nature of self-assessment in the daily practice of healthcare professionals, and a failure to properly explicate the role of self-assessment in a self-regulating profession’’. Members of the group felt that future researchers would do well to consider the relevant literatures summarised in their article (Eva and Regehr 2005) before attempting to articulate their own research questions.

Nevertheless, quantitative comparisons of assessment accuracy continue and need to give the time and resource invested in self-assessment through healthcare education and training. The concept of self-assessment has to be tied to its manifestation in practice. One of the intents and outcomes of the development of the suite of assessment tools that would be employed in the Foundation ePortfolio was that they would be observed and measured for their effectiveness. Further to that, an examination of the extent to which self-assessment was ‘effectively conceptualised … in daily practice’ was examined in detail with the extensive data in the case study of Chapter 4.

Future research could shift the focus to individuals' cognitions about their own developing clinical competence. This might, for example, explore the kinds of cognitive pathways that underpin self-assessment and performance, to clarify the relationships between self-efficacy, self-concept, motivation, self-assessment, and performance (perceived and externally measured). Qualitative research on the influences on the judgements that people make about themselves, the effect of interactions with and feedback from peers on self-assessment, and the triggers in everyday practice that highlight learning needs would provide a platform of information on which to build. Where there is doubt about the effectiveness of self-assessment interventions, randomized controlled trials could then be constructed on a well-defined theoretical basis, to determine their effect on the accuracy of determination of learning needs, or on subsequent learning activity and change in clinical practice. Current appraisal systems and the increasing use of multi-source feedback in the health professions lend themselves to research of this nature, and could be usefully informed by such research.

2.12.1 Informing the Next Steps

This review identified and substantiated the evidence for the effectiveness of self-assessment, and highlighted the opportunity to test the three core questions with a year's data from medical trainees. It provided a sensible template for the categorisation of the trainees in the case study (Chapter 5) according to their self-assessment behaviour. It also informed the core components of the case study's primary research tool (the ePortfolio).

The case study involved a cohort of medical trainees that provided the environment to evaluate whether the self-assessment review's findings (such as peer and faculty assessment as being more accurate than self-assessment, and poor performers being equally poor at self-assessment) could be replicated in a large year-long section of educational activity. Where possible, the findings of this first review would inform and be tested by Chapter 5, as would the areas that were identified as having insufficient evidence.

Self-assessment frequently is recorded within, if not enabled by, an (e)portfolio, yet the effectiveness of these tools had not been held to rigorous examination. The following chapter, the second systematic review, explores the effectiveness of portfolios as an assessment medium.

Summary Points

- The review's original questions were unable to be fully answered, largely due to a paucity of evidence of sufficient quality.
- Peer assessment is far more accurate than self-assessment, and it is better aligned with faculty/supervisor assessment.
- Competent practitioners are the best able to self-assess; the least competent are the least able to self-assess.
- Practical skills may be better assessed than knowledge or "soft skills".
- Feedback and benchmarking can play a useful role in improving self-assessment.
- There is no conclusive evidence that gender is related to self-assessment ability.
- Few studies explore the acceptability of self-assessment as a method, or the conditions under which it is taken.

3 PORTFOLIO SYSTEMATIC REVIEW

3.1 BACKGROUND

Like self-assessment, the use of portfolios in postgraduate health education has grown rapidly in the last number of years without a comprehensive synthesis of the evidence to their effectiveness. Portfolios are now used extensively for a disparate range of tasks, critically for educational progression and certification of summative assessment (including self-assessment), but also reflective practice, professional organisation, and learning. Their use has been promoted by institutions and regulatory bodies – such as the Royal Medical Colleges, the Nursing and Midwifery Council, Modernising Medical Careers and the Postgraduate Medical Education Training Board.

The author's institution of employment, NHS Education for Scotland, is an organisation that had promoted the use of ePortfolio. As the self-assessment review was concluding, the authorbegan to examine the evidence base for portfolios in postgraduate settings as well. At its broadest this was an examination of portfolios' educational effectiveness; but specifically relevant to this thesis it was an examination of whether self-assessment can be supported by portfolios. The findings of this review would go on to inform the design of the case study, as well as the NHS ePortfolio itself. This portfolio systematic review drew upon the methods and experience of the self-assessment review; it was entirely conducted by staff of NHS Education for Scotland, and the authorinitiated the review, collated the team, nominated the lead, undertook the literature review, adapted the coding sheet and was a principal author.

Traditionally portfolios have been artistic (and then financial) compilations of

documents for presentation, but more recently the term has come to encompass the collection, management and presentation of a far greater diversity of material for an increasing array of professions. But as portfolios in healthcare education are now used for a range of purposes, including delivering summative assessment, supporting reflective practice, and aiding knowledge management processes. They are seen as a key connection between learning at organisational and individual levels. With portfolios' migration to the electronic medium the extent and depth of their usage continues to grow as they, for example, integrate with e-learning platforms and enable rapid analysis of data supporting learning.

Amongst the healthcare professions, nursing has a history of using portfolios for reflective practice and they are now required by the UK Nursing and Midwifery Council. But recent years have seen portfolios contributing to educational provision under the auspices of many regulatory bodies and professional organisations. For example, in the UK in the field of medicine they are used by some medical schools and following the introduction of Modernising Medical Careers, required by the Postgraduate Medical Education and Training Board, medical schools and numerous Royal Colleges of Medicine.

Crucially, the expanding and broadening use of portfolios in postgraduate healthcare education is being actively considered or used for training, recertification/revalidation and continuing professional development (GMC, 2012). For high stakes decisions in any setting, there is a clear need for validated assessment criteria against which to evaluate portfolio data (Tillema & Smith 2007), and as records and vehicles for self-assessment, portfolios were judged to be an ideal medium for evaluation.

Alongside the rapid growth of portfolio usage has been corresponding publication of a diverse range of evidence and descriptions of the work; however, much of this is descriptive and there has been little attempt to aggregate or synthesise high quality findings. Initial scoping work in 2005 established that no single study had comprehensively combined all evidence regarding the effectiveness of portfolio use. This systematic review draws together the evidence across postgraduate healthcare education and examines the implications of portfolios migrating from paper to an

electronic medium, building on Challis's 1999 guide.

The review also examined an aspect of rapid change in the use of portfolios – the transition between paper and electronic versions, the latter of which provides new opportunities for compiling and collating data in ways that were very difficult or impossible with paper. This was of key interest to NHS Education for Scotland, which piloted the first e-portfolio for Foundation medicine alongside a paper copy. This e-portfolio would come to replace paper across UK Foundation and beyond (as discussed in the next chapter), and this portfolio review heavily informed the decision making that partially enabled this rapid expansion.

The review commenced in November 2005, with the comprehensive search conducted in January 2006 and results from an update search in October 2007. Analysis was completed in December 2007, the paper written in 2008 and published in May 2009. The ongoing work informed the analysis of both the case study that carried the first systematic review's questions forward, as well as the development of the ePortfolio itself.

The research questions were therefore more broadly focused to ensure the time spent on the project reflected the wider interests and needs of the organisation. A final decision to concentrate on postgraduate evidence was taken given the extent of published evidence and the fact that a second review group, interested in the subject, had been formed. This second group, from the University of Birmingham, looked at the effectiveness of portfolios in undergraduate health education settings (Section 3.14).

3.2 AIMS

The review aimed to answer three research questions in order to meet a number of objectives:

1. Are portfolios effective and practical instruments for post-graduate healthcare education?
    - establish how effective portfolios are as instruments to support reflective practice
    - summarise the strengths and weaknesses of portfolios for conducting

formative and summative assessment, including self/peer/supervisor assessment

- synthesise the evidence on portfolio usage in the work place and how they can further education
- ascertain whether portfolios can accurately support the educational needs of learners

2. What is the evidence that portfolios are equally useful across health professions, and can they be used to promote inter-disciplinary learning?
    - determine any differences in the effectiveness of portfolio usage across the professions, and
    - reveal how they can be used to support inter-professional education
3. What are the advantages and disadvantages in moving to an electronic format for portfolios?
    - examine the impact and implications of migrating from paper to electronic format

The terms “effective” and “practical” were extensively considered against the broad experience of portfolios, and for the purposes of the review are defined as follows: An effective portfolio is one which meets the needs of the users, supports them to achieve the aim of the portfolio and delivers the required elements to an appropriate standard. A practical portfolio is one which is user-friendly, efficient in terms of the overall cost and time demands on both the user and the support team who maintain it.

3.3 Literature Search

The literature search was conducted across a wide range of sources relevant to professional education. The database search covered all relevant health as well as educational databases, and included: MEDLINE, British Education Index, ERIC, HMIC, EMBASE, CINAHL, British Nursing Index, TIMELIT and AMED.

The strategies were designed for high sensitivity to minimise the risk of missing potentially relevant articles. The search ran from the earliest available date in each

database (e.g. 1966 MEDLINE) to January 2006 and did not limit by language, geography, or research methodology. An update search was conducted in October 2007 to include evidence published during the course of the group's first wave of analysis. The full Medline search strategy can be found in Appendix 1. Additional strategies for the other databases were based upon this search using consistent syntax and terminology.

One member of the team conducted an initial filter of titles for clear irrelevance to the review, and then a list of titles and abstracts were distributed (where available) to randomly selected (and shuffled) pairs of team members. Reviewers read the available information on each citation independently and decided whether the full text should be ordered for appraisal. They compared their decisions and discussed anomalies, requesting the article if one or both reviewers were unsure.

Once reading full articles, the team were also asked to identify cited references that might be of importance to the review. A cited reference search was conducted in late 2007 on the highest rated articles and where appropriate these were obtained.

3.3.1 Grey Literature

Grey literature (evidence not formally or commercially published) searches are expected of all systematic reviews as by definition they must include all relevant evidence regardless of whether it is available in peer-reviewed/commercial databases; however, in practice the extent to which grey literature will be relevant is highly dependent on the topic under review. Given the portfolio review was likely to have relevant evidence such as internal university papers (which had been recently made accessible due to indexing improvements at Google), the author organised a substantial grey search.

On an agreed date in September 2007 and then again in November 2007 three of the team independently searched Google (UK) for grey literature. A variety of search terms were used, related to the effectiveness of portfolio usage for education or learning (Table 3).

There is no method to exhaustively search the entire internet. For grey literature

searches they practically conclude when the searcher is no longer turning up new relevant items. In this review the three individuals agreed to stop searching on both dates when no new items were retrieved for a period of 30 minutes. On the first date this "data saturation" equivalent was achieved between 90 and 180 minutes, depending on the individual. On the second date very few new items (which all turned out to be indexed/spidered since the first search) were revealed, and none of the three searchers each concluded in under one hour.

Table 3. Combinations of Core Search Terms Used

Term 1	Term 2	Term 3
portfolio e-portfolio personal development plan	healthcare health professional learning	research evaluation effectiveness

Each of the three team members reviewed retrieved citations for relevance to the review questions, and saved any potentially useful documents to a shared storage space, thus avoiding duplication. Each person committed two to three hours to this search; the second date ensured results were as close to saturation point as reasonably possible.

3.4 Selection of Articles

3.4.1 Inclusion and Exclusion Criteria

In order to conduct a thorough and pragmatic review of the literature; broad criteria were set (

Table 4). All study design types were included, as it was established by early scoping searches that in this field there was little experimental research. Letters, editorials and conference abstracts were obtained in case they referred to other work which may have provided some evidence.

Table 4. Inclusion and Exclusion Criteria for studies from search results

Inclusion Criteria	Exclusion Criteria
Research Questions 1 & 2	
Articles which, both: were about the use of a portfolio by a qualified professional group (in a healthcare setting) in an educational / learning / professional development context **AND** described one or more of the following concepts: - what you do with portfolios - what you learn by using them - how a portfolio is used - perceptions of effectiveness of portfolio usage (even if descriptive) - informal evaluations i.e. perceptions, thoughts, views of users or others? - formal evaluation of portfolio as tool - portfolios contribution to career development	Articles including only undergraduate students (see question 2 exception) articles where the portfolio was no more than a log-book or checklist of procedures or items
Research Question 3	
Articles which described any aspect of the use of an electronic portfolio.	Articles where the portfolio was no more than a log-book or checklist of procedures or items articles which only described the technical specification or implementation of a portfolio articles where the portfolio was not used for learning e.g. as a teacher's planning tool / or for collation of pupil's work
Article Types Included - All Questions	
- any publicly distributed document (to include published and listed in a literature database, published in a print or electronic journal, or a publicly available website) - any language (identifiable by English-language index terms) - any country of origin	

3.4.2 Types of Portfolio

The review group discussed the boundaries and grey areas of what constituted a relevant portfolio during the early phase of the review. The type of portfolio of interest would include a collection of information to facilitate learning, and indicate engagement with the portfolio by the user, above and beyond a list of items; e.g.

clinical procedures undertaken by the user. A precise definition was not pursued, as it was feared it may limit the generalisability of the review. Each article was considered on its own description of the tool used, how it was used, and was included if enough information was provided to distinguish the interactive learning or reflection element which was of interest to this review. This meant that the same term e.g. log book, may appear in one article representing a simple checklist tool (and thus be excluded) but in another it may incorporate a reflective element in which case the article would be included.

3.4.3 Types of Participant

The main focus of the review was on articles involving postgraduate healthcare professionals; this was agreed in collaboration with another BEME systematic review group based at the University of Birmingham (Buckley et.al, 2009) who were reviewing the literature to report on the effects of portfolio use on undergraduate student learning. The term "post-graduate" was defined as having graduated and is practicing as a professional, i.e. when an individual is employable in their field. It should be noted that Foundation doctors are not fully registered with the GMC until the successful completion of their first year. Outwith the UK, and across the health professions, however there are variations in the terminology for the status of an individual with a healthcare qualification or degree.

With regard to answering the question on electronic portfolios, an initial scoping search revealed little evidence. As this was an area of particular and growing interest, inclusion criteria were widened to include participants of all types (i.e. including teachers and students in all learning settings) for this part of the review. This constitutes an area of overlap with the Birmingham review.

3.4.4 Types of Outcome Measure

Evidence on any reported outcome measure that addressed the research questions was included. Anticipated categories of outcomes which would inform on the

effectiveness and practicality of portfolios in learning included:

- skill (e.g. communication, clinical examination, reflection / self-awareness (There is some debate over whether reflection / self-awareness should be considered as “skills”).
- attitude (e.g. views of learning and teaching, self-confidence, satisfaction);
- behaviour (e.g. level of portfolio usage, participation in further learning);
- efficiency (e.g. time taken to prepare portfolio).

Articles providing only procedural details of a portfolio implementation process rather than describing the learning involved were not included, as were articles which described only a portfolio product specification.

3.5 Assessment & Appraisal of the Evidence - Online Form

An online form was developed to store citation information and details of critical appraisal and data abstraction by each reviewer. This was of considerable benefit as the team were based in four locations across Scotland, and therefore it was desirable to agree standardised formats for evaluating and managing information. This also facilitated the process of data checking and analysis. A software programmer was recruited to develop the form to the team’s specifications; this was done as an ASP coded web application which stored form data in a Microsoft SQL Server 2000 database. Web access allowed users the ability to enter or check data at any internet-linked computer. Data was ultimately downloaded into another application (Microsoft Access) for synthesis and analysis.

Individual usernames were issued to the team, and everyone tested the system on several articles to identify technical bugs or elements which could be improved. A record was then created for every full-text article, and a link was made to a pair of reviewers so that they could click on it, and begin entering data when ready (more details below).

3.6 Evidence Appraisal - All Full-text Articles

The processes involved in the appraisal of the evidence were heavily informed by the experience gained in the preceding review. With a different topic and review group, there were alterations, but these were minor and largely superficial and the portfolio group came to accept what had worked well for this self-assessment one. Firstly the whole team read and scored five articles and discussed them in depth. This process allowed a common understanding of the elements required to achieve an acceptable standard for inclusion to be reached. These elements included study design (sample size and selection), execution of research elements, analysis and clear / fair reporting of results. The team preferred this method to a rigid points-based checklist to deal with the anticipated variety of study types. A quality score was applied on a scale of one to five: 1 (very low), 2 (low), 3 (reasonable), 4 (high) and 5 (very high) and the team established a good level of consistency. These terms are used throughout the rest of this review to indicate the score applied to cited studies. For example, a study with a random selection of participants, achieving a representative sample of a population (if clearly stated e.g. including baseline characteristics) would score as 4 (high) or 5 (very high) depending on its size. A study using a convenience sample, or a sample whose characteristics were not described, would not score higher than a 3 (reasonable). The process revealed that an additional level of refinement was required as some of the literature fell within the inclusion criteria but could not directly answer the questions (listed in section 1.2). Therefore a score for relevance to the research questions was added (on the same 1 to 5 scale). It was agreed that a minimum score (for relevance plus quality) of 7 out of 10 would be acceptable, but with a minimum of 3 on both measures (i.e. a score of 5 + 2 was not acceptable).

Each full text article identified by the literature searches was randomly distributed to two of the team, who read it in full, blinded, to identify whether it met the inclusion criteria, and to score its methodological quality. This data was entered into their own record for that article on the online form. The pair then discussed each score and their reasoning for any discrepancies. If these could not be resolved to mutual satisfaction during this discussion the article was referred to a third party within the team. This

happened on three occasions, and in one case the article was shared with the entire team to agree an appropriate decision. Pairs were shuffled, so each reviewer was paired with everyone else on the team during the review.

3.6.1 Critical Appraisal & Data Abstraction - Included Articles

Once the pair agreed that an article met minimum standards (i.e. scored 7 or more), it was assigned to one of them to fully appraise, and extract data which answered one of more of the research questions. The team member paired with them for scoring was available to check or clarify any issues, but as little complex data was retrieved, double extraction (pairs of reviewers doing the task separately then comparing the results) was not undertaken.

The online form comprised a detailed checklist for appraising different types of research method or analysis employed (including literature reviews). For every full text article, assigned reviewers were asked to:

- rate the appropriateness of the article design to answer their research questions;
- describe the design and methodology;
- rate how well the study was conducted;
- rate the quality of the analysis and reporting;
- record the main findings and conclusions
- assess the study's impact level (Table 5); and
- note any issues or concerns they had about the study quality or relevance to the review.

3.6.2 Study Impact Level

Kirkpatrick's hierarchy is used when reviewing evidence to indicate the extent to which a study reveals the impact of an intervention on participants (Hutchinson, 1999). For example, a survey of users may report on their interaction or involvement with the portfolio, demonstrating a level one impact, in that they are engaged with the

intervention. A before and after study may show that users' attitudes or knowledge level were changed by the portfolio (level two impact) or that users incorporated learning into their work (level three). A more detailed description of Kirkpatrick's hierarchy adapted for medical education by the BEME collaboration group is given below.

Table 5. Kirkpatrick's (1967) Hierarchy Adapted for Medical Education by BEME Review Groups

Level	Description
1	Participation – covers learners' views on the learning experience, its organisation, presentation, content, teaching methods, and aspects of the instructional organisation, materials, quality of instruction.
2	Modification of attitudes / perceptions – outcomes relate to changes in the reciprocal attitudes or perceptions between participant groups toward intervention / simulation. Modification of knowledge / skills – for knowledge, this relates to the acquisition of concepts, procedures and principles; for skills this relates to the acquisition of thinking / problem-solving, psychomotor and social skills.
3	Behavioural change – documents the transfer of learning to the workplace or willingness of learners to apply new knowledge and skills.
4	Change in organisational practice – wider changes in the organisational delivery of care, attributable to an educational programme. Benefits to patient / clients – any improvement in the health and well-being of patients / clients as a direct result of an educational programme.

As mentioned above, the Kirkpatrick hierarchy was employed in various BEME systematic review groups, often with modifications to match the particular review questions.

3.6.3 Methods

The studies identified had insufficient homogenous or quantitative data to allow meta-analysis or formal synthesis. Reviewers individually identified all pertinent themes arising from each included article's findings. The evidence base was then discussed in its entirety and themes were collated into related groups according to how they meaningfully answer or inform this review's research questions. These grouped themes form the structure of the results section in the form of a detailed narrative

description of the evidence.

3.7 RESULTS

From the main electronic database searches 376 articles were found to meet this review's inclusion criteria. These were independently scored by pairs for quality and relevance to the review questions; 46 met minimum standards and were included.

After approximately eight hours spent on the grey literature search (Section 3.3.1), an ultimate saturation point was not reached, but it was agreed it was impractical and not productive to keep searching as no new results were turning up. Forty six articles were identified of which four met the inclusion criteria.

Citation follow-up and expert contact provided a further six articles which met the threshold, as did 46 from database searches. Therefore 56 articles were included in total (Figure 4).

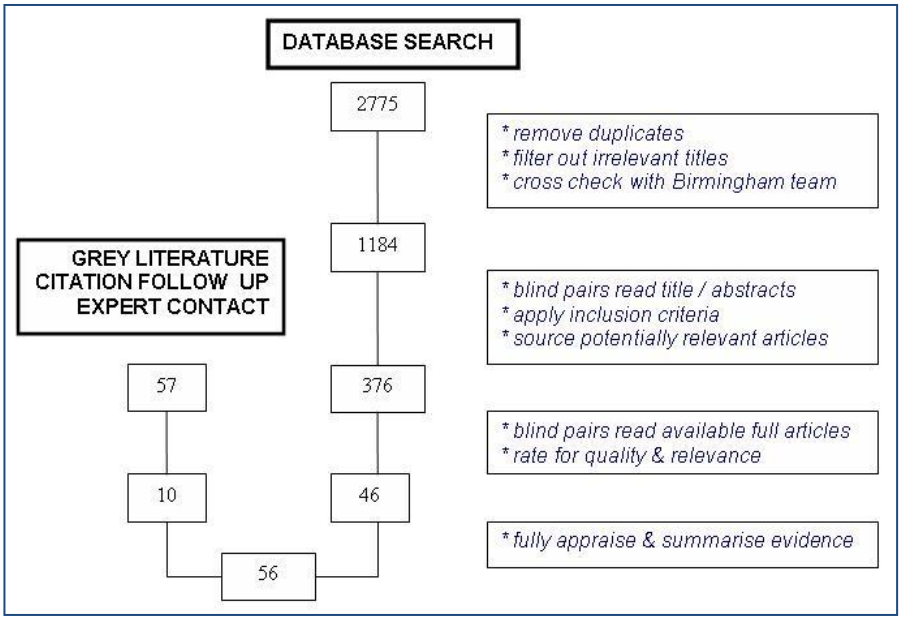

Figure 4 Flowchart of Search and Selection Process Showing Number of Included Articles Identified at Each Stage of the Review

3.7.1 Geographic Distribution of Articles

Included studies were conducted in ten countries (Figure 5). Almost half of studies were conducted in the UK (46%) and almost a third from the USA (29%). There were four each from Canada and the Netherlands, and one each from six further countries. The dominant majority of papers originating in the United Kingdom is notable and attributable to the early adoption and proliferation of portfolios in this country (search did not limit by language or geography).

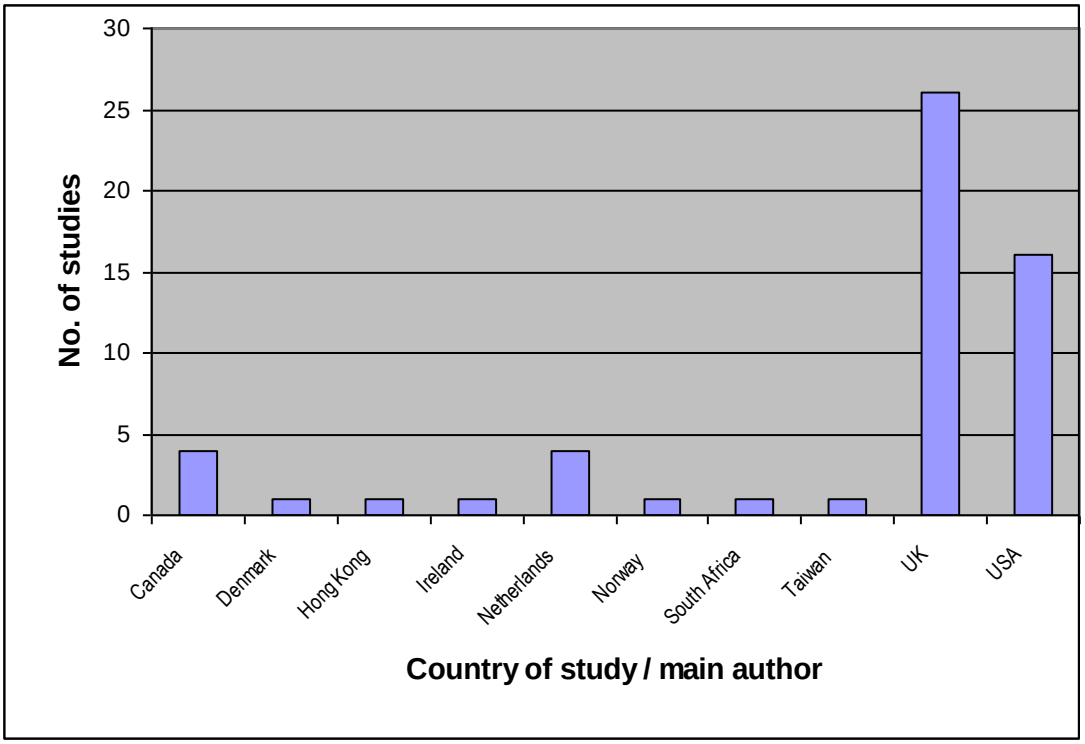

Figure 5. Location of Included Portfolio Studies (or Main Author if not Clearly Stated)

3.7.2 Professional Group Participating in Included Articles

Among the 56 included articles, seven different healthcare professional groups were represented, most commonly medicine (n=27) and nursing (n=12) (Figure 6). Of the articles in medicine with a clearly stated setting, thirteen were based in hospitals and ten in general practice. Other groups of postgraduate portfolio users included trainees in counselling and educational technology.

Undergraduate students (included only for the electronic portfolio question) were predominantly medical and teaching students, and 'other' groups included school teachers, principals, and educational supervisors.

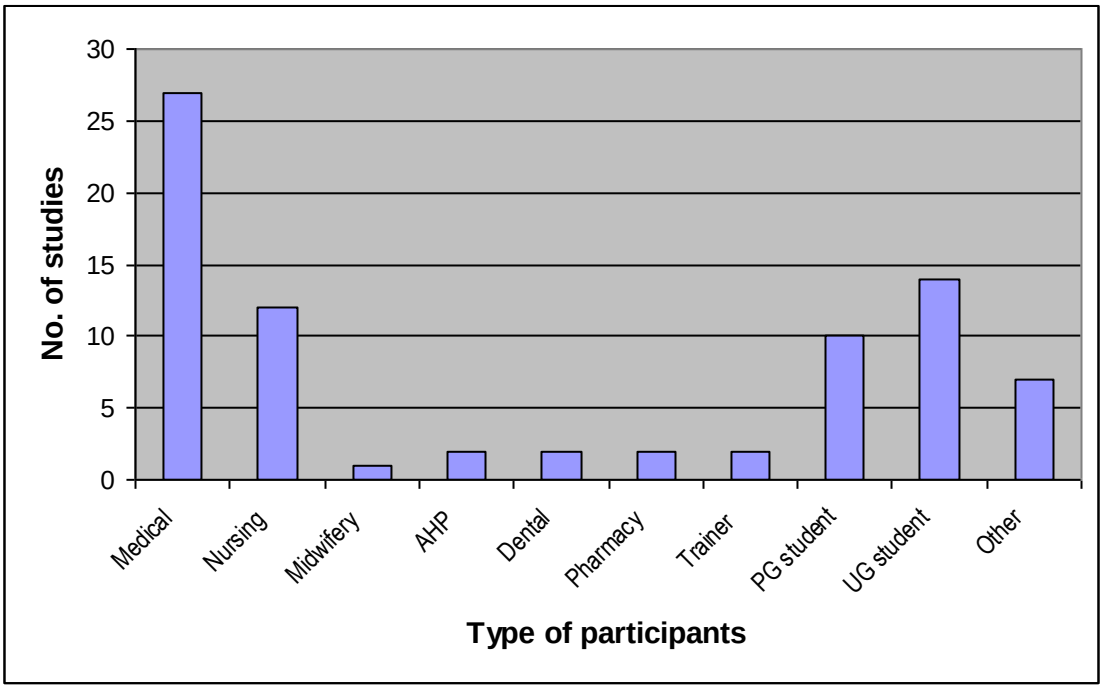

Figure 6. Professional Groups Involved in Included Studies (UG Students & Non-Healthcare Setting Participants Included in 'Other' Relevant to Question 2 - Electronic Portfolio Only)

3.7.3 Description of Included Studies

On the basis of study design, execution and reporting more than half of the included articles just exceeded the quality threshold scoring 3 out of 5, and were therefore defined as "reasonable" quality (n=32). Twenty four scored 4 (rated "high" quality). None were rated 5, i.e. "very high" quality.

The most common study design was uncontrolled observational (n=33) (see Figure 7). There were also ten comparative studies (six observational and four experimental) and six literature reviews (three of which were described as systematic reviews). This categorisation was not always straightforward as some articles did not follow a recognisable methodology, or did not report it clearly (seven remained uncategorised -

primarily descriptive reports).

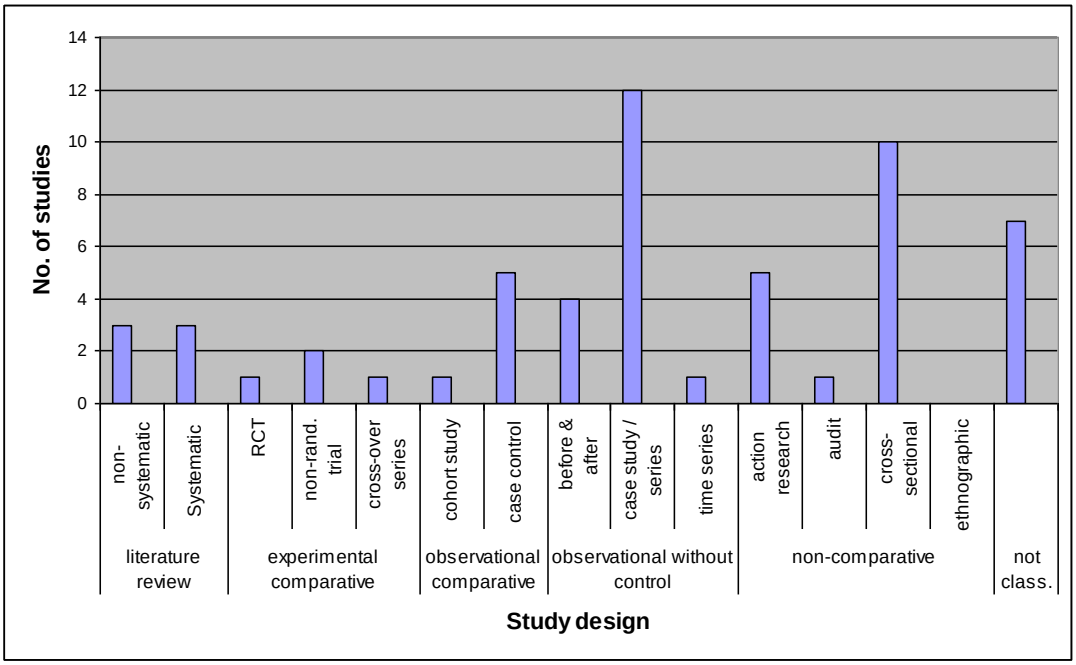

Figure 7. Comparison of Study Designs and Types of Included Articles by Number

The range of portfolio type used was very broad, and this review included all which involved the key element of user reflection or interaction with the contained information, for example a portfolio attempting to link learning to professional recertification, through to a very different one used to develop a counselling case profile. In many cases, descriptions of the content of the portfolio were scarce, therefore taking generalisable messages from the evidence base was not straightforward or justified.

According to Kirkpatrick's Hierarchy, most included studies were found to impact on the learning of the portfolio user (a level 2 impact, n=26), with fewer demonstrating effects on behaviour (level 3, n=10) (Figure 8). Two were found to indicate some effect on organisational change or benefit beyond the portfolio user (level 4).

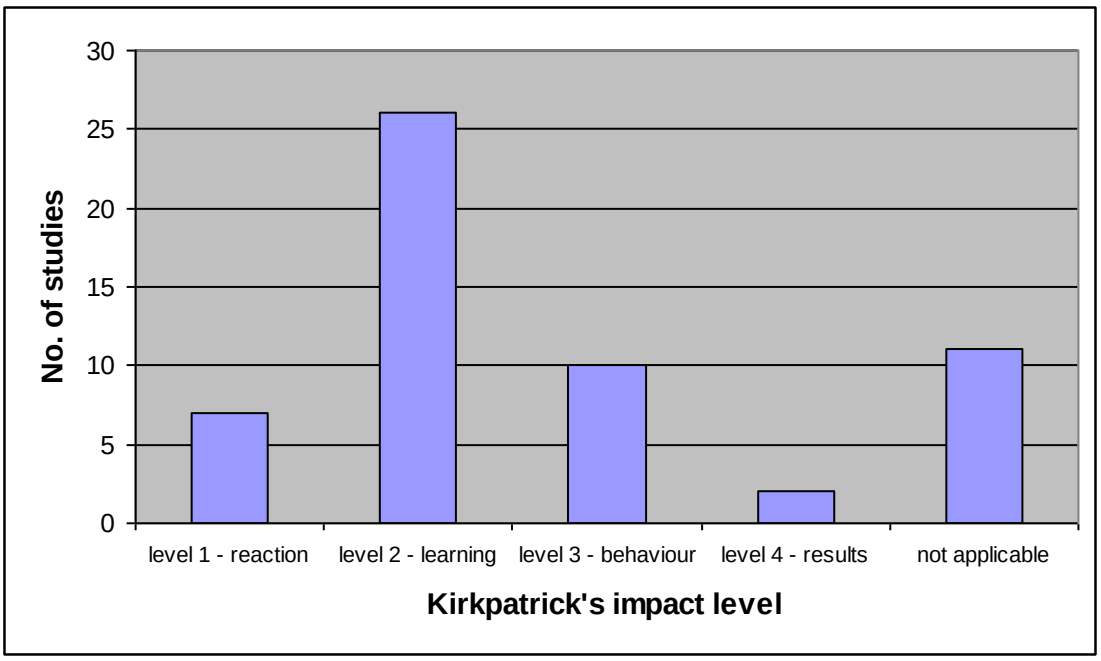

Figure 8. Kirkpatrick's Impact Level of Included Studies by Number of Studies

3.8 Are Portfolios Effective and Practical Instruments for Post-Graduate Healthcare Education?

This section reports relevant results from all 56 articles which met the minimum quality threshold. Under each theme, evidence from every relevant included article is presented. For the six included literature reviews which were found to meet the minimum standards for quality and relevance, evidence of relevance to the review questions and populations of interest are reported followed by additional primary evidence identified by the review. Higher quality studies (i.e. scoring 4 rather than 3) are given prominence in each section.

The review identified 38 articles which describe or test various aspects of the effectiveness and practicality of portfolio use. The evidence is grouped under the following themes:

- factors influencing portfolio use;
- use of portfolios for assessment;
- outcomes of portfolio use.

3.8.1 Factors Influencing Portfolio Use

The evidence brought together in this section demonstrates the extent to which the effectiveness and practicality of a portfolio (to an individual or an organisation) are influenced by a range of factors. These include users' positive or negative attitudes; gender; different levels of organisational support during implementation; early or sustained support / mentoring and the challenges of the time and cost involved in portfolio use. This section examines the evidence for factors influencing use in general, obtained from 23 relevant articles, but where authors specifically examine the electronic medium, or compare electronic with traditional portfolios, the topic is discussed in the later electronic section.

3.8.1.1 User Attitude

A UK study of personal development plan (PDP) usage in general practice (GP) medicine reported somewhat contradictory attitudes in users (Cross & White, 2004). Whilst 64% of respondents (n=277 in total) reported submitting their PDP as a means to obtain Post-Graduate Educational Allowance (PGEA) accreditation and 53% agreed a PDP was a "*hoop-jumping*" exercise, their attitudes to the educational value of PDPs were simultaneously quite positive – depending on the educational tool. Only 42% found the portfolio (referred to as a "regional workbook") of use and 36% valued SWOT analysis; however, 61% valued the use of Patient Unmet Needs and Doctor's Educational Needs (self-directed learning tools), 74% valued the reflective practice and 81% thought the Significant Event Analysis component was valuable. These survey results, based on a strong postal response rate (81%), convey wide variation in what general practitioners value in their PDP with the high rating of some tools seeming to contradict the notion that the PDP is merely a form-filling exercise. The potential cynicism expressed by many in completing PDPs was also balanced by the fact that 82% of respondents saw the PDP as forming a substantial part of their revalidation.

A small, well-conducted two-part study (focus-group, semi-structured interview) of UK general dental practitioners also reported that portfolios could be well received in revalidation in this sector (Maidment, et al., 2006). Feedback from the volunteer group

was largely very positive about the potential for revalidation. They felt that including a system for appraisal would be beneficial, although the small (n=10) size of this study within primary care dentistry may limit the generalisability of the findings.

3.8.1.2 Gender

Murray's 2006 UK study used quantitative and qualitative analysis of e-portfolio data from Pre-Registration House Officers (PRHOs), their educational supervisors, nurses, nurse supervisors and two cohorts of further education (FE) students (the latter group outwith the inclusion criteria of this review question), about engaging with portfolios (grey literature, 2007). The authors compared portfolio use by gender, and showed that a greater proportion of female users accessed the portfolio following training (64% vs. 55%), but were less likely to progress from being a 'reader' to a 'poster'. Once using the portfolio, females were more likely to remain users and qualitative analysis indicated that they were more likely to perceive and describe positive educational effects. However, these analyses did not incorporate the effect of being a voluntary or mandatory user.

3.8.1.3 Implementation Method

Low initial compliance rates were reported by a USA surgical study, after implementing their Surgical Learning and Instructional Portfolio, a case-based portfolio that included self-assessment and reflection (Webb *et al.*, 2006). Although the programme director and coordinator actively tried to improve compliance, the rate remained <50% and residents (n=40 in total, but early numbers are not clearly reported) did not rate it highly. No detail was provided of the implementation process to this point. The processes were revised in 2004 to include monthly feedback, topic collation and coded discussion as new resources, e-mail contact with the supervisor and quarterly notification of incompletion to all relevant parties. Once put into practice, the lessons learned from the initial implementation saw compliance rise to 100% and considerable higher appreciation from residents. The article cited "*dedicated faculty review*" and "*perceived importance of the project*" as critical factors in successful implementation.

This article would have benefited from the provision of more detail, particularly on the initial implementation work, but does provide reasonable evidence on how embedded communication and feedback during implementation can influence uptake.

Other articles reported similar limited evidence from doctors in other specialties. Snadden & Thomas (1998), conducted a qualitative action research study in a geographically diverse area across the north of Scotland on portfolios in GP vocational training. This revealed doubts regarding the introduction of portfolio learning without "*intensive support at a one to one level*". Their work, which included extensive interviews and focus groups with 20 pairs of trainers and their trainees (four were unavailable and one pair refused to participate), concluded the implementation process for portfolios might be more important than the structure of the portfolio itself. In 2006, Kjaer et al. reported on the implementation of an e-portfolio for GP medicine in Denmark (n=90 GP trainees). Similarly, this article did not set out to measure the implementation process, but cited proper time and scheduling, consideration and provision of information about the portfolios use to users, and a "*practical technical demonstration*" as being key to proper implementation.

Murray's UK study demonstrated that from a user's initial contact with a portfolio to their full engagement with it, the key factor in uptake is its relevance to the individual (Murray, grey lit 2007). As previously mentioned, in this study the portfolio was used by PRHOs, nurses, and two cohorts of further education (FE) students. Use was compulsory for PRHOs (n=33) and voluntary for other users (n=171) and this is reflected the proportion who accessed it (88% vs. 55%) and made entries (88% vs. 23%).

3.8.1.4 Mentoring / Support

The impact of constructive interaction with a mentor or supervisor on portfolio use has been explored in a number of studies. Driessen, et al., (2007b) state in their recent review of the effectiveness of portfolios in medicine, (30 included articles, of which nine were in the postgraduate sector, five Continuing Medical Education) that mentoring made an important contribution to the success of the portfolio, but a

definition of this success was not clear.

The following evidence describes the influence of mentoring on the process of portfolio use, but less on how mentoring affects outcomes. Considering the initial uptake of portfolios among potential learners, Webb, *et al.*, (2006) found that compliance among surgeons in training increased from less than 50% to 100% when, as previously described, monthly feedback from a dedicated supervisor was introduced. Snadden & Thomas (1998), in a qualitative study of portfolio use among 44 trainees in general practice, reported that the portfolio was '*usually not adopted where there was no support from the trainer'* or where tensions existed in the trainee / trainer relationship. This was illustrated by means of a few case studies which did not explore possible confounding variables. Pearson & Heywood (2004) achieved a good response rate (77%) of registrars in a UK deanery when evaluating a pilot portfolio for 92 GPs. Authors reported that users with a supportive trainer more commonly used their portfolio for reflection on their practice.

Few studies looked at the potential impact of mentoring on sustained portfolio use, but Snadden & Thomas (1998) found that the majority of their study group had stopped using the portfolio by months six to eight of the training year, '*despite the intense effort to support portfolios in the region'* (1998). In his study looking at uptake and subsequent level of use of electronic portfolios among cohorts of PRHOs, nursing students (under- and postgraduate) and sixth form school pupils, Murray (grey literature, 2007) found a relationship between the provision of feedback on the portfolio from a mentor and the frequency and level of use by the learner. A comparison of 46 learners who received feedback with 22 who did not, showed that 57% of those who received feedback went on to become classified by the author as '*continuous users*' versus 0% of those who had not received feedback on their initial postings. However, it is not reported which of the cohorts these learners were from and this finding should be interpreted with caution as the terms of use and purpose of the portfolio were very different for each cohort. Likewise, the timescale of the project was unclear, so that the term 'continuous' does not give any indication of actual duration of sustained use.

There is some suggestion in the literature that, for some individuals, mentor support was needed for reflection. The assessors in a previously cited study of dentists (Maidment, et al., 2006), expressed this opinion, although the dentists themselves had mixed views. Tiwari & Tang (2003) made the observation that some of their learners (twelve postgraduate nurses in Hong Kong) appeared to lack sufficient cognitive and reflective skills to make best use of the portfolio. They recommended that support be tailored according to need.

Users have also reported concerns regarding supervisors with insufficient knowledge or understanding of the portfolio. Ryland, *et al*. (2006)conducted a pilot study into portfolio use amongst second year Foundation doctors (i.e. doctors in the first two years of postgraduate training) in the UK (n=147) in 2005/2006. Using qualitative analysis of free text questionnaire responses (response rate: 65%), the article stated one of two emergent themes as educational supervisors "*needed more guidance on how to use the portfolio*". Although the study was relatively simple, the deanery that conducted it used the evidence the basis for the roll-out of consequent supervisor training as they believed there was a "*continuing need to emphasise the educational value of the portfolio by both Foundation trainees and their educational supervisors*". Hrisos, et al., (2008) in a UK study noted that over half of Foundation trainees (n=182) felt their educational supervisors (n=108) were not sufficiently knowledgeable about the portfolio.

Lack of support was identified as a factor which was considered to limit the potential of the portfolio from a survey involving 121 nurses in the UK (Richardson, 1998) and in another survey of 90 GP trainees (Kjaer, et al., 2006). One outcome of focus groups conducted by Chabeli (2002) with 20 postgraduate nursing students in South Africa required to complete a portfolio for a semester for assessment purposes was, that they felt that teachers should, *'constantly monitor and provide support and guidance to the learners during the preparation and compilation of the portfolio'*. Similarly, Coffey's (2005) in a survey of nurses (n=22) using a portfolio for assessment for a diploma in gerontological nursing in Ireland, found that respondents felt more support was needed in completing the portfolio. It was implied that mentoring should be the

vehicle for this support. McMullan, *et al.* concluded in their 2003 review of the use of portfolios in the assessment of learning and competence for nursing, that it was important for the tutor to provide regular support and feedback, *'as this helps them build their portfolio'*, likewise, Bowers and Jinks (2004) reported (from a limited evidence base) that UK practitioners needed guidance and support.

3.8.1.5 Peer Support

A small number of studies explored the influence of peer support on portfolio users. Mathers et al.,(1999) conducted a crossover study comparing traditional and portfolio method of PGEA, and used a model of three facilitated meetings of groups of UK GPs (n=32) compiling portfolios for PGEA purposes during a six-month study period. Authors reported that this process provided a supportive stimulus to learning and was an appropriate use of time by the GPs. A survey conducted by Austin, et al., (2005) of 1,415 Canadian pharmacists highlighted the value of an information-sharing session, allowing participants to discuss experiences with colleagues in a facilitated environment. It was reported that after this session, the feedback from subjects indicated that they were, '*far more informed, aware and supportive of the portfolio concept*". In Tiwari & Tang's (2003) small study of nursing students, portfolio users spontaneously developed collaborative learning strategies and gave each other support, apparently as a result of being involved in the portfolio process.

3.8.1.6 Time

Many authors reported time as a factor that had a negative influence on portfolio use by healthcare practitioners (e.g. Keim, et al.,, 2001; Dornan, Carroll & Parboosingh, 2002; Maidment, et al.,, 2006; Jensen & Saylor, 1994; Dagley & Berrington 2005; Duque, *et al*. 2006), as they had difficulty adding portfolio use to their already busy schedules. Kjaer, et al., (2006) had doubts that the 10-15 minutes allocated protected time could be worked into the existing trainee / trainer interaction. In the GP PDP study, Cross & White (2004) reported 73% (of 204) respondents as "disagreeing" or "strongly disagreeing" they had enough protected or unprotected time to undertake

their PDP. Seventy four percent of this group also “agreed” or “strongly agreed” that the PDP study competed with the time reserved for their socialising and family. No studies objectively tested the implication that time was a barrier to the practicality of portfolio usage.

Mathers, *et al*.’s (1999) crossover study cited above, demonstrated that portfolios take a considerable and very varied amount of time (at least for new users). The time involved in preparing a portfolio for PGEA was 24.5 ± 12 hours (range 10.5 to 64 hours): much more than the fifteen hours which could be claimed for the process. The implications of this additional time on the relative efficiency or “*amount of educational gain*” of traditional pattern of PGEA was discussed by authors as it does not allow for practical elements e.g. travel time to courses, preparation and follow up. Authors report that these issues make comparison with the portfolio approach more equivocal.

Keim, et al., (2001) showed that dietetics professionals assigned to use a portfolio (n=661) conducted learning needs assessments significantly quicker than a control group (n=714) following the traditional route (2.7 ± 2.6 hours vs. 4.4 ± 5.1 hours, p=0.002). They were also quicker in developing learning plans (4.0 ± 4.9 hours vs. 2.4 ± 1.9 hours p=0.018).

3.8.1.7 Cost

Although studies allude to savings made by adopting portfolios (particularly electronic versions) such as reduced administration cost or printing, a single small study substantiated the claim. Moyer (2002) reported feedback from four of thirteen nurses who used a portfolio in the USA, and compared the traditional cost of nurse credentialing (>$40,000 per examination) with the cost of portfolio evaluation of the same content ($14,752). Among the retrieved articles there were none examining finance and its potential influence on individuals’ portfolio use. However, note that the review did not search specifically for economic articles or have cost-effectiveness as part of the inclusion criteria, therefore the author does not draw further conclusions.

3.9 Use of Portfolios for Assessment

Twenty two articles reported on the use of portfolios around the assessment of healthcare professionals at work exploring the ways in which they have been used for formative or summative types of assessment, and exploring the boundaries of reliability and validity.

3.9.1 Reliability Summative Assessment

Several articles reported on the reliability (ie a measure of consistency and accuracy) of using portfolio assessment for summative decisions about healthcare professionals – sometimes referred to as "high stakes decisions". Six articles examined by Driessen, *et al.* (2007b) in their systematic review of portfolios in medical education, gave an 'average' reliability of 0.63, although the range of scores of the six studies cited was unclear. Increasing the number of raters raised the reliability towards a value of 0.8 as usually required for high stakes decisions (by regulatory bodies, educational panels, etc.). Also reported were a number of measures which had positive impact on inter-rater agreement i.e. training, rater discussion, global criteria with rubrics. Lynch (2004), whose literature review included portfolio assessment as part of a wider focus on practice based learning for residents and physicians, and who cited similar articles to Driessen, reported a slightly more negative view. A key focus was on studies by Pitts, *et al.* (2002) who looked at portfolio assessment with 8 GP trainers. They achieved poor to moderate inter-rater reliability of 0.1 to 0.41 which increased to 0.5 with rater criteria discussion. McCready (2007) carried out a literature review on portfolios as a tool for assessing competence in nursing and also reported the literature as ambiguous with regard to reliability (n= 15 included studies). She questioned whether conventional tests of reliability and validity can be brought to bear on the holistic data presented in portfolios (referring to Pitts, *et al*, 2002). The literature review by McMullan, *et al.* (2003) focussed on the use of portfolios in nursing and concluded that there were difficulties in assessing portfolios using purely quantitative methods.

3.9.2 Enhancing Reliability

As already highlighted, Driessen, *et al.* (2007a) reported some successful strategies to improve reliability; use of small groups of trained assessors and discussion amongst raters before (and sometimes after) the assessment. These findings were supported in McCready's literature review. Jasper & Fulton (2005), although reporting on the development of marking criteria for practice-based portfolios, tested their new criteria on 30 portfolios at two UK sites where Masters courses in nursing and healthcare disciplines were offered. They concluded that the use of double marking with an external examiner along with explicit descriptive criteria against which portfolio content could be judged, was the way forward.

Alternative strategies to improve reliability were raised by other authors. Melville, *et al.* (2004) reported ratings of all paediatric Specialist Registrars' (SpRs) portfolios in one UK deanery (n=76). In the first year portfolios were assessed by a single rater, and the following year by two raters. They concluded that although their method of portfolio assessment could not be used as a single assessment method for high stakes decisions, without multiple observers (assessors) or observations, it had a place as part of a triangulation process with other assessment methods. In two studies identified in McCready's review, tri-partite meetings during the portfolio assessment process were used. In the first study this tri-partite assessment was between the academic supervisor, practice mentor and subject (post-registration nurse). It reported the subjects as having valued this approach (n=15, 75% participants). The other article, although there was little detail provided, suggested the tri-partite meeting was crucial. Another study by Jarvis, *et al.* (2004) looked at portfolio entries representing thirteen psychiatric skills from eighteen psychiatry residents in the USA. A total of 80 entries were examined in the light of the six ACGME (Accreditation Council for Graduate Medical Education) general competencies. They found five out of six competencies represented in the portfolio and similarly concluded that whilst it was desirable for a single evaluation method to assess competencies, it was reasonable and realistic to use more than one form of evaluation to examine performance. Maidment, (2006) reported on a portfolio developed with a range of specific sections to meet dental

professional body requirements with regard to providing evidence of fitness to practise. Based on the study sample of 10 general dental practitioners, they concluded that when using the portfolios for revalidation the scheme would be significantly enhanced by using it as the basis for an appraisal interview, thus triangulating the data and its interpretation.

3.9.3 Validity for Assessment of Competence

The validity and reliability of portfolios assessment are often combined in the literature making it difficult to distil clear messages. There would however, seem to be tension between balancing both reliability and validity of portfolio assessment with learning.

On the positive side some studies found portfolio assessment valid for specific criteria. For example, in Mathers, *et al.* (1999) comparison of traditional route to PGEA accreditation with a portfolio based learning route for GPs, the breadth of topics covered in the portfolio was extremely wide and entries were seen to be appropriate for the claimed educational objectives. Jarvis *et al.* (2004) as described previously, examined portfolio entries in the light of the six ACGME general competencies. Although all general competencies bar one were represented, they concluded that all the competencies could be covered with some revision of the portfolio guidelines. O'Sullivan (2004) tested the reliability and validity of eighteen psychiatry residents' portfolios in the USA. Scores were compared with another cognitive performance measure and global faculty ratings on clinical performance. The author concluded that portfolios provided valid evidence of competency although the evidence was not strong.

Other authors expressed more uncertainty or concerns. Smith & Tillema (2001) looked at portfolio use in the Netherlands among different types of professionals and in different settings which included senior nurses (unit leaders, n=26) and nursing staff (n=33. Interviewees (n=12 unit leaders) highlighted the perception that the evidence found in the portfolios was considered to have questionable validity, especially when it is used for assessment and is no longer a working portfolio: *'if the evidence is original,*

who chooses it and what is the quality of the various portfolio entries?' The literature review by Carraccio & Englander (2004) focused on portfolio assessment in medicine and reported the difficulty in striking a balance between the creative, reflective aspects of the portfolio which is learner focused with a structure that is reliable and valid. Finally the small scale pilot by Maidment, (2006) found significant concerns about the use of a portfolio for revalidation to meet dental professional body requirements: *'revalidation* [using a portfolio] *doesn't prove you are a good or a safe dentist, it proves you can fill a book'.*

3.9.4 Linking Portfolio to Quality Assessment Frameworks

A small pilot study (Dagley & Berrington, 2005) evaluated the way in which a portfolio was used by UK GPs (n=5). This included logging critical incidents and attempting to link revalidation categories to elements of their PDP and CPD actions. These links, however, were found by the authors to include some inconsistencies, and they proposed this area required further training. PDPs were quality assured against two published CPD frameworks: Rughiani's, and the Cromarty Eastern Deanery matrix. They were found to have evidence of a continuous learning 'spiral' and to contain rich material. However audit, and the more objective elements were underused.

3.9.5 Compliance

It seems evident that when portfolios are required for summative assessment, compliance is greater. Driessen et. al., (2007b) noted that if portfolios were not formally assessed, their use tailed off (based on Pearson & Heywood, 2004; Snadden & Thomas, 1998). Smith & Tillema (2001) inferred from user comments that because keeping a portfolio was not required, participants did not find time for it in their daily work. McMullan, , *et al.*(2003) also identified a study in which participants were less likely to use the portfolio if assessment was not present although no data on this was presented.

However, the point was also confirmed by Murray (grey literature, 2007) as previously

noted in his study of implementing e-portfolios with mostly healthcare professionals in four colleges. He found that after training for all, only 23% (n=171) of users who were given the choice of using the portfolio (for others it was compulsory), actually used the system.

3.9.6 Formative Assessment

Reviews by both McMullan, *et al.* (2003) and Kjaer, et al.,(2006) found there was considerable support for portfolios to be used for formative assessment. Kjaer et al. (2006) carried out a study with a cohort of GP trainees and an on-line portfolio (n=79 portfolio users, 11 non users) and used two evaluation questionnaires (one for users, one for non-users) which had been validated for construct and content validity and which collected both quantitative and qualitative data. They found that the portfolio was a good basis for formative assessment and recommended that a part of the portfolio should be kept exclusively for formative feedback. Although not distinguishing between formative and summative assessment, the article by Tiwari & Tang (2003) reported on the qualitative data collected through semi structured interviews with twelve of the study participants, selected according to criteria specified in the article. They found that portfolio assessment can have a positive effect on learning and users reported a distinct preference for the portfolio form of assessment over the standard approach (written assignment and end of term test). Webb, *et al.*, (2006) in a study with a cohort of surgical residents, concluded from the user survey (40 residents) that the most beneficial aspects of portfolios was the educational aspect e.g. the faculty interaction and feedback. Similarly a study by Coffey (2005) with 22 postgraduates from a nursing programme reported findings were mainly '*positive regarding the effect of the assessment on their learning*' and gave some quotes to back this finding. Finally Smith & Tillema (2001) identified the importance of feedback provided by the portfolio regardless of whether it was formative or summative – it gave an opportunity for subsequent improvement of actions.

3.9.7 Influence of Assessment on Portfolio Contents

Driessen, van Tartwijk *et al.* reported (based on two studies by Driessen (2005) (not included in the review) and Mathers, *et al*. (1999)) that there was no conflict between using the portfolio for summative assessment and learning in the postgraduate sector and that they can be successfully combined. However, there is some evidence to the contrary. McMullan, *et al.,* concluded in their literature review that portfolios become assessment led, resulting in a reduction in learning value. Three primarily qualitative studies also addressed the formative / assessment conflict. Snadden, *et al.* (1996) through an action research project with 20 pairs of GP trainers and registrars, reported that participants perceived that formal assessment would inhibit the type of material collected in the portfolio, but it must be noted that these perceptions were not substantiated by any differences in portfolio content. In the latter part of the Webb, *et al.* (2006) study, when 40 surgical residents (100%) complied with the use of the portfolio, only 20% felt that their portfolio should be used for resident assessment although no reasons were given. Kjaer, (2006) with 56 (71%) of portfolio users showed that GP trainees feared they would be less honest and avoid showing shortcomings, if their notes were used for assessment purposes. On a similar point Murray found that assessment impacted on the type of engagement displayed by the users: 55% of assessed users only submitted entries to the required sections compared to 41% who used it continuously.

3.10 Outcomes of Portfolio Use

Many articles alluded to outcomes of portfolio use, however, as will be discussed in more detail later, most failed to clearly or objectively demonstrate that self-reported or measurable effects are in fact due to portfolio usage. The following sections describe some evidence from seventeen articles which did attempt to demonstrate true outcomes.

3.10.1 Promoting Reflection

The encouragement of reflection is a commonly cited purpose of a portfolio, and there is some evidence that this is facilitated by portfolio use. In one study, simply providing a portfolio appeared to have a positive impact on users' attitudes to completing activities that were previously unsupported by portfolios. Keim, et al., (2001), randomly assigned dietetics professionals to either a portfolio group or a control group who reported Continuing Professional Educational activities in the traditional format (Cronbach's alpha > 0.75). At the two-year follow-up (79% response rate from 1,082 surveys), a significantly greater proportion of the portfolio group (79% vs. 46%) reported that they had completed considerably more self-reflective entries in the previous 12 months ($p<0.001$). A five-year evaluation of portfolio use by six to ten surgeons per year (total n=40) indicated that 72% of users felt the portfolio should be used for self-reflection (Webb, *et al.*, 2006). This contrasts to 42% of GP trainees in a study by Pearson & Heywood, who actually reported using their portfolio for reflection (2004); and 56% of educational supervisors who felt their trainees were encouraged to reflect by use of a portfolio (Hrisos, et al., (2008)).

Other authors have reported some adverse effects. Swallow, et al.,*et al.* (2006) found some negative views among 25 UK community pharmacists, some of whom felt that the portfolio could actually inhibit reflection if there was a lack of confidence about how the information may be used "*against them"*, a view echoed by Pearson & Heywood (2004). Austin, et al., (2005), pointed out that some users already described themselves as being reflective, and believed that being forced to use a tool for this purpose interfered with their own approach to their professional development.

Some authors state that users can use portfolios to reflect but few describe how reflection is defined or measured making it difficult to determine whether it is a meaningful outcome or one which has a knock on effect on professional practice. Dagley & Berrington (2005) found that some records showed evidence of users completing a reflective cycle - this was shown by electronic links with recorded incidents from their practice, their PDPs and CPD activities. The two-part study by Maidment, et al., (2006) also reported on the potential for portfolios to support

reflective practice. Among the ten participants however, there were reports that reflective practice took place regardless and therefore portfolios were an artificial and unnecessary imposition. The concept of the portfolio as a "*burden*" was also raised in Hrisos, et al., study cited above, with two thirds of the trainees reported that the collection of required paperwork was difficult to manage in busy hospital wards (Hrisos, et al., 2008).

3.10.2 Learning / Knowledge

Tiwari & Tang's (2003) controlled study is probably more usefully considered as a case study, as the two groups of users are at different stages of learning – with the control arm being undergraduate students following traditional assessment methods, whereas the group of interest to this review were postgraduate nurses using a portfolio. Ten of the twelve participants interviewed reported positive academic effects of the portfolio, including a deeper understanding of study topics, and the process of learning itself. The attitudes of users were cited as explanation though as the remaining two participants were reported to be '*only interested in getting a degree'*. Webb *et al.* found that 75% of users (30 of 40) felt that the portfolio had improved their understanding of a topic they were studying.

Coleman, et al., (2006) conducted a controlled study in the USA using two cohorts of graduate multicultural counsellors (n=28) who were assigned to use a portfolio or case formulation method to demonstrate their competence. The final exams were rated blind to group allocations, and showed a significant difference with the case formulation group rated higher than the portfolio group. The lack of detail on participant characteristics and randomisation procedure for the study however, makes this comparison somewhat unsafe. There was a high inter-rater agreement (0.67-0.79) on the quality of portfolio contents.

3.10.3 Engagement with Learning

Mathers, *et al*., (1999) crossover study of GPs using traditional or portfolio PGEA

methods undertook an experimental study design but presented the analysis in a qualitative narrative style, not taking into account any effect of the crossover itself on outcomes. It states that there was evidence of completion of a learning cycle by portfolio users who reported a mean number of seven (± 4 SD) critical incidents which subsequently modified their learning objectives i.e. evidence that portfolio caused people to adopt principles into practice more than PGEA route. The method of analysis and reporting unfortunately mean it is not possible to determine when the effects happened in relation to the method being used by the user at that point; any lasting effect beyond the six month period on each approach and any effect of which came first.

Keim, et al.,(2001) showed that, compared to control, their portfolio group produced more learning needs assessments (71%-22%, $p<0.001$), and more learning plans (70%-12%, $p<0.001$). Overall though, measures such as attitude towards professional development, self-efficacy to conduct a learning needs assessment were reasonably positive at baseline but did not change significantly by two-year follow up (paired t-tests, $p>0.05$) The perception that portfolio maintains competence was not rated positively by either group and again did not change significantly between baseline and follow up. In Mathers, *et al*. (1999) portfolio users were found to tackle a much wider breadth of learning activities and study topics.

Fung, *et al*. (2000) conducted a multi-centred non-randomised trial in Canadian obstetrics and gynaecology departments, giving an advanced year of exposure to a prototype portfolio (described in Walker, et al., 1997*)* to residents at one school, and then a comparison of measures with three other schools as they embark on usage of the full internet-linked version. Compared to control, the residents using KOALA (Computerised Obstetrics & Gynaecology Automated Learning Analysis) reported increased awareness of their self-directed learning ($p<0.05$), were more inclined to learn on their own ($p<0.015$), had a positive attitude toward life-long learning ($p<0.000$) and expressed strong interest in taking on new learning ($p<0.018$). This well-cited study also reports the impact on their perceptions of their future learning. They felt a clinical experience portfolio would now contribute to their residency ($p<0.011$)

and that didactic lectures would not be sufficient to support their future learning ($p<0.028$). This study was limited however, by a number of factors, including lack of information about group comparability at baseline, insufficient detail on the timing of data collection and the fact that the intervention consisted of a year's exposure to the portfolio's prototype. Although authors concluded that the internet-linked portfolio has positive effects, it may have been the advance year of the (non-internet linked) prototype which had these effects.

In Keim, et al., (2001), both the portfolio and control groups demonstrated generally positive attitudes towards assessing learning needs and developing learning plans across the two-year follow up: ratings showed no significant difference between groups (t-tests). Both groups were slightly less positive however that the portfolio maintained competence (scores around 52-54, on a scale where the midline is 55). Tiwari & Tang (2003) found that all twelve portfolio users reported a high level of satisfaction in using the portfolio, once the initial lack of confidence about the process was dealt with. They expressed pleasure in the freedom afforded by this method of assessment.

The evaluation (n=147) conducted by Ryland, *et al.* (2006) concluded that a portfolio (used by UK Foundation doctors) did support educational processes; trainees reported positively on the role of the portfolio in supporting assessments and enhancing reflective practice. The size and response rate of the study were relatively low, however, and the study was reported in brief.

3.10.4 Supporting Learning into Practice

Coffey (2005) evaluated a clinical learning portfolio for gerontological nursing by means of a postal survey of the programme's first graduates. The author reported an unexpected and tangible result in that respondents' use of the portfolio continued on to their subsequent clinical practice. However, the study had inherent weaknesses, including a small sample (n=22) of a single cohort, the survey instrument not being tested for reliability and validity, and there was no description of the qualitative analysis used. In Austin et al., (2005) study of 1,415 pharmacists using a portfolio in

Canada, users completed a mean of 5.6 learning objectives per year (range of 0-10). Almost two thirds of self-identified learning objectives were achieved (63% ± 25%) which resulted in a mean of 2.2 changes to practice, facilitated by the portfolio. Campbell, *et al*. (1996) found that two thirds of study participants (n=152 Canadian physicians) reported that portfolio use made them reflect on patient care, and to take note of which educational activities enhanced their expertise.

3.11 Are Portfolios Equally Useful Across Health Professions; Can They Be Used To Promote Inter-Disciplinary Learning?

No evidence was identified to allow us to answer this question - a small number of studies were found which included for example nursing and midwifery, or postgraduate and undergraduate medical students, but no sub-group analysis was conducted to allow understanding of the relative needs of the different groups or the different ways in which they engaged with the portfolio.

It is likely that this reflects the traditional divisions between the healthcare professions where each works independently from undergraduate level through to continuing professional development. Although some organisations are beginning to promote multi-disciplinary learning it may be some time before the commonalities between the professions are recorded in any standardised or comparable way using e-portfolios.

3.12 What Are The Advantages and Disadvantages in Moving to an Electronic Format for Portfolios?

The team identified nineteen articles which provided evidence on this question. Note that as electronic portfolios were of special interest to the review, wider inclusion criteria were adopted, to include undergraduate students and articles conducted in a non-healthcare setting.

The main messages extracted from the evidence were grouped under the following themes:

- factors influencing e-portfolio use
- outcomes of e-portfolio use.

A variety of factors were seen to influence the usage of e-portfolios and the significant ones are listed below.

3.12.1 Electronic Medium

One good quality study directly tested the effect of the electronic format on portfolio use. Driessen, et al., (2007a) conducted a randomised trial of two types of portfolio format with year one medical students in Maastricht. Five of seventeen mentors were randomly selected to participate (all agreed) and the two groups of students each was responsible for, were randomly allocated to either paper (n=47) or web (n=45) based portfolio. Although the comparativeness of groups was not described, it is assumed that the (unspecified) randomisation procedure adequately minimised bias. Pairs of raters independently scored the portfolio content for quality of evidence and reflection (coefficients 0.71-0.91). The scores were very similar with the notable exception of the 'additional effort' of the web-based population with the perceived effort they applied to creating their portfolios. This manifested in more personal approaches to the look and content of students work. There was strong evidence that the medium of the portfolio influences the amount of time users are willing to spend with it. There was a moderate effect size of 0.46 indicating that the web group spent more time on developing their portfolio (15.4 vs. 12.2 hours; p=0.05). Both groups were similarly satisfied with their portfolio. The article's discussion refuted the notion that extra time was required for the web versions, and hypothesised the electronic medium motivated the users to spend more time with the portfolio. There was unanimous agreement from mentors (n=5) that web-portfolios are easier to use as they allow faster retrieval of evidence through hyperlinks, and enabled access from a variety of sites at the mentor's convenience.

Chang (2001) conducted an evaluation of an electronic portfolio used by (an unspecified number of) undergraduate teachers assessing its functions and impact on students' educational progress. Most respondents felt it was beneficial to use the

electronic medium to access others' e-portfolios. A finding also described by Clegg et al., (2005). The vast majority (93%) of Chang's students believed they could improve the standard of their own work by having the option to view their peers'. Students found the feedback from peers more helpful than that of their instructors, which authors speculate may be due to higher expectations of instructors and demands on their time to provide extensive information. There was 80% agreement that using peers' portfolios enhances communication with those peers. The electronic medium therefore enabled sharing and exchange of information that would not be possible in paper format.

Fung *et al.*, 2000, is an often quoted study cited as demonstrating the positive effect of the electronic medium, however as previously mentioned it appears that the comparisons made are between residents at one school exposed to a prototype e-portfolio for a year ahead of three other schools who all then used an internet-linked version of the same tool. The additional positive learning effects may therefore be attributable to the advanced exposure to the tool rather than the electronic medium.

Banister, et al., (2006) highlighted the importance of piloting new e-portfolio systems, in their study which revealed that an in-house system was better suited to their purpose (teacher education in the USA) than a commercially available one. This is echoed by Scott & Howes who reported learning important lessons about improvements required to the interface of a new portfolio system, following a pilot with UK medical students (grey literature, 2007).

3.12.2 Data Transfer / Accuracy Across Systems

A portfolio's ability to support an individual's life-long learning necessitates the transfer of the relevant records and information through one's educational and professional transitions. In theory, the electronic medium would be an ideal medium to ensure one could have continuous access to all relevant past items. In reality, Horner, et al.,*et al.* (grey literature 2007) in a series of case studies illustrated the difficulty in transporting data between different e-portfolio systems in further and higher education institutions across England. Concerns regarding the security or

confidentiality of data contained within electronic portfolios emerge in many studies (for example Carney & Jay 2002).

Dorn & Sabol (2006) demonstrated in a multi-site before and after study conducted in the USA, that paired rating scores correlated well for artistic portfolios assessed in both paper and digital formats. Assessment scores for the digital portfolios were slightly higher than those on paper, but were a good predictor (significant at alpha 0.05 level, confidence interval 0.96 - 1.03).

3.12.3 Users' IT Experience / Skill

Students' experience in information technology correlates positively with their perceptions of learning through an electronic medium and therefore, use of the portfolio model. Hauge (2006) measured this in their Norwegian interview of five student teachers and survey of 76 students (beta = 0.38 $p<0.05$). Dornan, *et al*. (2003) conducted a qualitative case study which describes the evaluation of a web-based portfolio, demonstrating that students appreciated the design, for example, the ease of navigation.

Kjaer, et al., (2006) developed and validated a questionnaire to evaluate the use of a new online portfolio by 90 Danish GP trainees (79 of whom had used the portfolio and eleven had not). The response rate was over 70% for both groups. Whilst two fifths of respondents (39%, n=56) stated that they would not have started using an e-portfolio if given the choice, after the study, 87% agreed that they preferred the electronic medium. With regard to post-study use, 50% agreed that they would continue using the portfolio the same amount, and 46% expected to increase their use. Some portfolio users were wary of the perceived potential for external control of their learning. It was described as being more appropriate for formative than summative assessment, in that it could be used as a prompt for discussion points with a trainer.

Whilst the electronic medium requires support and training especially for those less familiar with the technology, any portfolio system would require this from an educational perspective. "*It is frustrating when the trainers are not completely familiar with the use of the portfolio. The time spent with the trainer should be used to discuss*

educational issues – not technical issues" (Kjaer, et al.,2006). Trainees noted that the hospital setting may make the use of an electronic portfolio problematic (with access to computers) unless a PDA version was available. Non-users of the portfolio related common responses to why they felt unable to use the portfolio including: lack of information; protected time and support from trainers; access to ICT and personal motivation.

3.12.4 Training and Support for e-Portfolios

The training and support that users receive was frequently cited as a factor that influences their uptake of portfolios. Redish, et al., (2006) in their description of the migration of a paper to web based portfolio in a graduate education programme, exemplify what many articles relate by concluding, "*training for both faculty and students is critical to successful implementation and ongoing technical support should be given careful consideration*". Unfortunately they do not substantiate this sentiment by linking it to research.

Similar to the other factors influencing portfolio use, training and support were not directly evaluated as an intervention in most studies. Duque, *et al*. (2006) provide the single instance the author found of evaluation of training against a control in this Canadian study of 133 medical trainees on a geriatric rotation, though they do not measure the training's influence directly against usage. The study evaluated students use of an e-portfolio divided into control (no training) and intervention (introductory hands-on session) groups, surveying both students and tutors. Students' comfort with the e-portfolio was surveyed immediately post rotation and at the conclusion of the clerkship year (response rates 98% and 55%). The first survey revealed 66% felt they "strongly / somewhat" agreed they felt comfortable, compared to 48% of the control ($p<0.05$). The survey at the end of the clerkship year found that the difference between the groups comfort levels had disappeared, following a significant increase in the control group and decrease in the training group (both $p<0.04$) (final scores: 57% and 56%). Tutors in the Duque *et al.* study were surveyed once, and were asked to rate training as a limiting factor in use of the e-portfolio. None saw it as a strong limitation,

30% as moderate and 60% saw training as having no limitation on their e-portfolio use. Support was viewed in a largely similar way with the helpdesk availability seen as strongly limiting by 10%, moderately by 20% and of no limitation by 40%. From these results it would appear that most of these tutors did not regard training and support as significant factors influencing use, but the size of the sample (n=18) and (critically) the fact the results were not measured against actual usage by the tutors, would call into question how much the tutors' results should be generalised.

3.12.5 Outcomes of e-Portfolio Use

Two significant outcomes of e-portfolio use were noted in the included literature: engagement with learning and turning learning into practice.

3.12.5.1 Engagement with Learning

The potential for the portfolio to capture the dynamic aspects of learning, particularly in relation to the student / tutor relationship was illustrated by Duque, *et al.* (2006). Their case control study of 133 undergraduate medical students found that the e-portfolio was perceived to be a more effective feedback tool than more traditional methods ($p<0.04$). These perceptions were given further weight by a demonstrable increase in the number of portfolio entries made by both students and tutors. Portfolio entries were only validated if they included comments and action plans, illustrating a quantifiable ongoing record of self-reflection with an average of 30 entries in one month. From this limited evidence they concluded that the inclusion of comments and action plans, and the engagement of both the student and the tutor in these evaluative entries showed that the portfolio was more than an information repository, but a dynamic account of learning, reflection and supervision.

Chang (2001) reported that a web-based portfolio was perceived to have had a positive impact on learning processes across a number of areas, with 47% of students "strongly" agreeing and 42% agreeing. These positive findings were echoed by Bartlett & Sherry (2006) on their USA study of 34 undergraduate and postgraduate teaching

students.

3.12.5.2 Learning into Practice

The potential of the portfolio to bridge the perceived gap between the curriculum and the individual learner, or between teaching and practice, was examined by a number of studies including Avraamidou & Zembal-Saul (2003) and Jensen & Saylor (1994). In Jensen & Saylor's study (may be n=49 but not clear) of physical therapy and nursing students in the USA. Students identified that the process of portfolio completion allowed them to structure their learning and reflection as well as place learning in the context of completed practice. The authors advised against measuring or assessing portfolios, stating that the aim of portfolios should be to inform, not to measure. They concludes that portfolios are '*more valuable for what they do than what they are*', suggesting (as Duque *et al.* 2006) that the very process of portfolio completion can be a learning experience but only with the support of mentors, tutors and the organisation as a whole. However, the evidence to support this conclusion was meagre.

Cotterill, *et al*.'s (grey literature 2007) study of electronic portfolio implementation in two UK medical schools highlighted the potential contribution portfolios can make to organisational practice. They contrasted experiences in introducing portfolios to undergraduate medical students in two medical schools using questionnaire feedback from around 500 students. Around 80% of students from one medical school thought that the portfolio was a useful learning experience, and as well as helping students plan and organise their learning there is some evidence that portfolio use prompted reflection (72% spent time reviewing what they had learned). However in the second medical school, only 39% reported that recording their learning helped them to think about the process of learning. The portfolio appeared to be perceived as somewhat separate to the 'real work' of the curriculum, indicating that perceptions of the role and purpose of the portfolio may affect the ability of students to engage fully in portfolio use to develop learning. Swallow, *et al.* (2006) showed that portfolio use was beneficial in the planning and organisation of nine UK pharmacists' professional

activities. Although these studies were in the undergraduate environment, they were included as there were no published postgraduate equivalents.

3.13 Discussion

This review takes a broad and pragmatic look at all types of evidence regarding the effectiveness of portfolios across post-graduate healthcare education (and beyond for electronic formats). While it is important not to lose sight of common sense when attempting to evaluate an evidence-base with potential recommendations for decision makers or practitioners (Smith & Pell, 2003), it is unavoidable to conclude that there remains a lack of objective examination of the effectiveness of portfolios. Although exploratory and uncontrolled investigations can be informative, there was a tendency towards reporting statements not backed up by evidence. The same unsubstantiated opinions of an author (or portfolio users and trainers) sometimes then repeated as fact in subsequent publications. This along with insufficient studies being conducted with due consideration of study size or sampling, failure to use an appropriate and clear intervention, no consideration or reporting of characteristics of participants and non-participants, make the body of evidence less than robust. With substantial funding going into widespread, and sometimes mandated, portfolio use, coupled with high expectations of what those portfolio systems can deliver, it would seem highly desirable that every opportunity be taken to properly investigate and test how portfolios are implemented, designed and supported allowing generalisable messages for other users and providers. Proportionate evaluation should be built in as a key feature of new portfolio projects, but research which generates generalisable findings will be of most value.

3.13.1 Portfolios: Practical Instrument for Education?

The evidence base contained many examples of portfolio in regular use by professional groups in the workplace across the healthcare and educational sectors. It was apparent that planned, supportive implementation of a portfolio was a vital step in enhancing its

uptake and use by the target group. Evidence from successful implementations have incorporated buy-in at an organisational or faculty level, perhaps to create a purposeful and clear driving force as users begin to invest time in the portfolio.

There was good evidence to indicate that the support of a well-informed mentor can be a crucial factor in the uptake of portfolios. There was also evidence to suggest that it can influence the extent of portfolio use, particularly when specific regular feedback was provided. However, even when this kind of input was present, it was not always sufficient to ensure long-term sustained portfolio use. Competing demands on time often intervened and portfolio learners reported needing more support from faculty.

Other factors have been demonstrated to influence whether uptake and use of portfolios is achieved, including the characteristics, attitudes, experience and learning preferences of the users, however this evidence is less substantial in some cases e.g. gender of user. Many others are alluded to in the evidence base, but have not been objectively tested: including the availability and flexibility of users' time, access to computers, relevance and quality of the individual constituent parts of a portfolio. Unfortunately, there is no substantiated evidence that specifically examines portfolios' attributes (components, functions, linkages, core purposes) against how well that portfolio is used. Measuring a portfolio's use by altering the attributes and features that comprise it would be a comparatively simple task, and one that could be done retrospectively.

The status of the portfolio - voluntary or mandatory - is a crucial defining feature which directly influences user attitude, uptake, and the amount of time they are willing to spend on it. Therefore it should also influence the way in which evaluations or research should be interpreted. Clearly if professional registration hinges on its completion, users will put in the time required for this even to their own personal cost. However, they are likely to report concerns about use of their data, its security and suspicions regarding the purpose of monitoring. There is evidence that users may be simultaneously cynical about the purpose of a portfolio, but positive about its potential to them individually - this conflicting feeling by users has to be managed. Unless compulsory or an embedded part of the organisation's ethos, there is likely to be an

uphill struggle to achieve compliance.

3.13.2 Portfolios: Effective Instrument for Education?

If well implemented, portfolios have been demonstrated to effectively further both personal and professional learning in a number of ways. There was evidence of increased responsibility for learning: i.e. portfolio users have been shown to be less passive about their own learning needs and plans for future learning (but without a baseline measure in most studies, this assertion is not robust). There is overall agreement that portfolios aid learning processes and outcomes. There are mixed views of whether portfolios aid or hinder reflection, with evidence on both sides - this may come down to the individual's learning preferences, or some aspect of the portfolio itself. Although some authors suggest that a mentor may be beneficial to support reflection, this hypothesis has not been directly tested. A small number of studies describe users' views of the benefits of peer support. These include a more positive attitude towards portfolios and as a stimulus for learning. But in virtually all studies a substantial minority of users fail to engage with the portfolio. No studies were found which thoroughly investigated reasons for non-compliance or resistance to portfolio use. Future research work on portfolios would benefit from taking these (and other) important confounding variables into account, and may allow refinement of successful portfolios already in use.

The outcomes occurring as a result of portfolio use are a direct way of assessing their effectiveness. However few articles were found which tested a meaningful control between, or within groups of users, or looked at a comparison intervention in order to reliably reveal outcomes of portfolio use. Many were cross-sectional or case studies of one particular portfolio, evaluating users and / or supervisors' feelings and experiences of the portfolio or the supporting processes after a fixed period of use. These articles were an often-quoted source of beneficial effects or positive reports of portfolio use in the literature. While looking at these provided an insight into the range of ways portfolios are used, and how successful they were individually, the generalisable messages were limited. These 'snapshots' of portfolio use failed to measure baseline

characteristics of users (or give any indication of characteristics of non-users), meaning that positive or negative outcomes were impossible to attribute confidently to the portfolio. Few made attempts to identify confounding variables and incorporate them into the presentation of results e.g. the level of experience with portfolio or self-directed learning, ability to use or access appropriate technology, attitudes to learning, learning style - which were all alluded to as reasons why a portfolio was or was not successful.

3.13.3 Portfolios for Assessment?

The meaningfulness of attempts to rate portfolios have been questioned in the literature, and there remains a lack of evidence in terms of inter-rater reliability. There was wide variation in published studies on the level of reliability of portfolios for summative assessment (principally conducted in medicine). It is clear that reliability increases with more raters or discussion between raters, but this incorporates additional time / cost, and it is unclear what size or direction of impact this would have on the ultimate scores. Evidence from both medicine and dentistry described the importance, to both practitioners and assessors, of triangulating portfolio data with other assessment methods.

Quantifying portfolio content and use may be too simplistic to capture professional learning and engagement and some authors reported that portfolios should not be used for summative judgements but instead for more qualitative and less structured personal development. It may be that more structured portfolios can and should be assessed, particularly for students and newly qualified professionals. However, as individuals progress through their career, qualitative methods of judging the portfolio may be more appropriate to allow the less tangible learning outcomes such as professional values and judgements to be captured. This depends on the type of portfolio, and attempting to generalise from a range of types may be unhelpful. These however lose the potential for individualised features which allow users to focus on developing their own needs and learning.

There was more positive, but weaker quality, evidence that portfolios are effective and

useful for formative assessment. However, to date this mainly comes from a theoretical understanding of the potential analysis of the information obtained within portfolios, rather than objective tests that this process works well or is meaningful.

3.13.4 Advantages and Disadvantages of the Electronic Format?

By definition a portfolio in the electronic medium offers the advantage of additional flexibility in a number of ways. This included flexibility of access to the information for users and supervisors, and virtually unlimited potential variation in content. This appeared to inspire or motivate users: good quality evidence was found to show that electronic portfolio users were willing to spend longer on it than those using paper-based portfolios, although ultimately self-reported satisfaction was similar between the two groups. A longer term analysis of these groups may be interesting to determine if the additional time spent provides a benefit. Ready access to peer's portfolio work was rated by some users as a particular advantage. The review found a small amount of good evidence that electronic portfolios were more effective than a direct comparator in paper format both as a feedback mechanism, and for encouraging reflection in users.

An electronic portfolio may be readily linked to competency or quality assurance frameworks, or to users' PDP / CPD activities. These links can be automated and updated far more simply in the electronic format. Such links, however, particularly with mandated portfolios and those used for sensitive assessments or high stakes decisions may trigger security concerns.

Many authors cite training as being important when implementing an electronic portfolio, and this is likely to be a requirement when implementing an electronic portfolio system, as there was evidence that users' technical ability and knowledge significantly affect how they interacted with it. Technophobia remained an issue for many users, and if portfolio content is to be assessed, users must be adequately equipped to enter appropriate information, and not disadvantaged by their lack of confidence. However few have investigated this: e.g. the frequency, duration, format

or content of training, to identify the key elements. The provision of technical support should be distinct from education support to contend with such issues.

There was reasonable evidence that moving from paper to electronic can be done accurately and that assessments of the same material in both formats are well correlated. The transferability of data between e-portfolio systems (required to facilitate life-long learning) is tentatively successful at the moment with some pilot projects now published but the process is far from straightforward.

True (and safe) interoperability has to be achieved before the full potential of e-portfolios to support lifelong learning is realised. Nevertheless the evidence indicated that progress was being made towards the realisation of standards that will sustain the transfer of data between e-portfolios.

This review was published as "The effectiveness of portfolios for post-graduate assessment and education: BEME Guide No 12" Tochel C, Haig A, Hesketh A, Cadzow A, Beggs K, Colthart I, and Peacock H. Medical Teacher, 2009;31(4)299-318.

3.14 Undergraduate (Birmingham) Systematic Review

At the same time as the postgraduate portfolio systematic review was being conducted, a separate BEME group (Buckley et. al., 2009) looked at the undergraduate evidence. Initially the groups had considered working together across both populations, but the large extent of evidence in both areas, geography and expertise on the post and under graduate sides being concentrated in one or the other (Birmingham University or NHS Education for Scotland) led to the decision to split the topic. Nevertheless, the two groups kept in regular contact throughout the process and shared ideas about methods and findings.

The Birmingham group also found the evidence base for their work was limited, but as

with the postgraduate review they were able to report on a number of areas, many of which were complimentary. Buckley et. al. found undergraduate portfolio users were more engaged with reflection, but the quality of these reflections was questioned by some authors. They also reported that portfolio usage improved with feedback. Similarly they found notable evidence citing the amount of time required of portfolio users and the impact it could have on clinical education – a tension well-mentioned throughout the literature.

The review also noted that a higher level of self-awareness was reported by undergraduate users, an issue not mentioned in the postgraduate evidence. Also in contrast to postgraduates, undergraduates using portfolios were reported to have a greater knowledge and understanding/knowledge of the subject matter but the evidence base was small and weak.

The Birmingham review did not report on generalisable portfolio issues (implementation, use of mentors, etc.) and because of this, and the fact it did not elucidate self-assessment or provide additional notable portfolio findings from the undergraduate population, it did not contribute significantly to this thesis.

3.15 Strengths of the Review

This BEME systematic review was based on a broad, sensitive search including all healthcare professional settings. All available articles were read, blind by two team members; non-English language articles were translated and a thorough grey literature search was undertaken. Good internal consistency was achieved for quality scoring and critical appraisal.

3.16 Limitations of the Review

The systematic review process was laborious and time-consuming, and proved very challenging for the team which was not based in an academic organisation. In the time taken to complete the work, another systematic review was published in the subject area, albeit with a narrower and exclusively medical focus.

While the online data entry form was extremely valuable and was of interest to other BEME groups, it is worth noting that considerably more time would have been required, in collaboration with the programmer, to develop it into a fully functional and user-friendly system.

3.17 Conclusions

Whilst there was an extensive and expanding evidence-base in this field, like the previous systematic review the heterogeneity of design and data, as well as questions around quality, made formal synthesis impossible. But also as in Chapter 2, the systematic evaluation of the evidence did, to varying degrees, inform various aspects about the use of portfolios in postgraduate healthcare education.

High level organisational support with a well-designed and sustained implementation was seen to be key to the uptake of portfolios. Mentors (supervisors) could have considerable impact on uptake as well, especially when regular feedback was given. Portfolios were revealed to be composite tools, and as such users could have complex or contradictory feelings about using them. There was also evidence, although somewhat limited, that portfolios helped learners engage. Summative assessment of portfolio content was also seen to be reliable between multiple raters – a point that will be tested in depth in the next chapter; similarly, the evidence suggested that triangulation with other sources was desirable which could readily be accommodated by an electronic portfolio, such as the NHS ePortfolio.

The opportunity to test the questions proposed in the evidence arose by using the NHS ePortfolio as a research tool of the case study in the next chapter. This was an unprecedented chance to test the empirical questions on a vast body of real trainee data. While this review identified the benefits of electronic portfolios over paper, the transition to electronic was already in progress within the health sector. The evidence base this review established was available to inform the development of the ePortfolio (a medium primarily designed to support assessment), and whilst this was possible in some instances the challenges involved in practice are fully revealed in this thesis' Discussion (Chapter 6).

3.18 Future Research

There are many gaps in the evidence, much of which appears to have been produced as a result of short-term local projects, e.g. rapid evaluations on specific portfolio projects. Several areas of research are urgently required to provide generalisable evidence:

- identifying genuine outcomes of portfolio use;
- identifying confounding variables underlying the variation in portfolio use among different learner types and professional groups;
- identifying the types of portfolio which are appropriate for the range of purposes they may be employed for: summative / formative assessment; creative / self-directed learning;
- assessing the cost-effectiveness of different approaches to portfolio implementation and the necessary support mechanisms;
- determining the differences in the effectiveness of portfolio across the professions, and revealing how they can be used to support education between the professions

Portfolios are increasingly expected to support education and training and many organisations, professional bodies and academic institutions are investing significant resource (financial and time) in introducing them to students, trainees and staff. Given the lack of high quality evidence, and gaps identified above, this may be premature. The ambitious and ever changing expectations attached to portfolios, particularly electronic portfolios, may risk losing sight of the fundamental purpose of the educational environment which portfolios were introduced to support. It is likely that the most appropriate portfolio to support summative assessment is different in nature and function to that best suited to self-directed learning. Anecdotal evidence may be useful to organisations selecting a portfolio to use, but a solid evidence-base relating to effectiveness, confounding variables, costs and outcomes would better support such decisions. Again, as with the initial review, this reinforced the author's view that a comprehensive evaluation of a year's training data would be a significant opportunity to confirm this review's findings as well as delve into areas where evidence was scarce.

Back in 1994 Jensen & Saylor stated "*we believe portfolios should be a recognized legitimate aspect of a course or program, not a busywork activity*" – a sentiment that has been echoed consistently by many authors since - both the belief in portfolios, and the concept of embedding them in study or work. It would appear that with substantial and sustained commitment at all levels when implementing a portfolio (organisational, faculty, mentor / peer / supervisor *and* user) it can facilitate a range of learning and work-based development.

Summary Points

- The quality of the evidence available precluded full answering of the initial research questions; there were very few objective evaluations of portfolio systems.
- Portfolio users were less passive than control groups.
- There was mixed opinion on whether portfolios aided reflection/formative assessment which may have been the result of different study conditions.
- The reliability of summative assessment scores within a portfolio improved with multiple raters and discussion between raters.
- Uptake of portfolio use was linked to its implementation and organisational support.
- Mentors influenced the extent and quality of portfolio use, particularly with feedback.
- The time required of all involved with a portfolio was frequently underestimated by stakeholders.
- The mandatory or voluntary status of portfolios had a critical influence over its use; reasons for non-compliance have not been thoroughly investigated.
- Users of electronic portfolios were more engaged than paper portfolio users, and there is a small amount of good quality evidence that the electronic medium encouraged reflection.

4 SCOTTISH FOUNDATION MEDICINE

The case study, which will be described in detail in Chapter 5, provided an opportunity to test the empirical questions from the first (self-assessment) review, informed by a greater understanding of the use of e-portfolios offered by the second review.

The Scottish Foundation ePortfolio contains an extensive dataset of trainee doctors' assessments by self, peers and supervisors. It offered real data which could be used to test the assumptions made in the literature and, where possible, evaluate areas where there has been a paucity of quality evidence. The ePortfolio for Scottish Foundation Medicine was already an established online assessment tool in 2007. Its component parts, as well as numerous annual revisions, made it an ideal environment for investigation as it operationalised assessment including (critically) self-assessment by combining the assessment processes in a single system.

The following section describes the structure of the Foundation Programme.

4.1 FOUNDATION MEDICINE

August 2005 saw the introduction of the Foundation Medicine Programme, a two year generic training programme that recognised the need to improve the early years of postgraduate training in the United Kingdom. Foundation aimed to implement the recommendation that "after graduating doctors should undertake an integrated, planned two-year Foundation programme of general training". The programme links undergraduate education to specialist and general practice medical training, in an outcome-based programme comprised of structured rotating posts designed to give experience in specific subject areas. A suite of assessment tools were designed to measure achievement in specific competencies. Although some of these competencies are generic, most Foundation posts are in acute care settings and the trainees learn to manage the care of the acutely ill. The Foundation programme aims to ensure that doctors are trained to a standard of general competence that ensures they are prepared for further specialist training and can deliver the highest standards of patient care.

Good Medical Practice was implemented in November 2006 by the General Medical Council and became the document and guidance for all doctors registered with the GMC. Underpinning the principles of the document is the notion of personal accountability where the practicing doctor "must always be prepared to justify (your) decisions and actions" and "recognise and work with the limits of your competence". Both concepts would be supported with effective self-assessment.

Foundation Year One is a transitional year where newly graduated medical students enter the National Health Service. Foundation Year One medics' GMC registration is only provisional and they must meet set specific requirements to successfully obtain their full GMC registration by year's end. Foundation Year Two emphasizes the care of the acutely ill, but also continues to build upon generic clinical skills from Year One, as well as softer skills such as time management, communication, and team working. In all trainees work in 65 specialties (e.g. paediatrics, haematology, infectious diseases) that all provide the opportunity to progress toward competence on specific "procedures", defined by the GMC as "demonstration of competence in a series of procedures in order for a provisionally registered doctor with a licence to practise to be eligible for full registration. These will be recorded and signed off within the e-portfolio. Evidence of completing these core procedures is also required for successful completion of the Foundation Programme".

Central to Foundation training is regular assessment based in the workplace. The assessments are core to providing public accountability for GMC registration and they document the development of the trainee as they progress through both years. Trainees are expected to perform below end of year competence until the latter stages of the year.

The Foundation Programme introduced a (paper) "learning portfolio" to manage its composite structure locations and supervision processes, as well as the documents within those two years. This includes the rotating posts (geographical and subject specific placements) and controlled progress through the posts.

The learning portfolio was comprised of five components shown in Table 6.

Table 6. Components of ePortfolio in 2007-08

	Component	Description
1	**Competencies**	the list of skills that are required to successfully complete the Foundation Programme
2	**Forms**	the documents required to record meetings with educational supervisor, reflective practice and self-appraisal
3	**Educational Agreement**	a document which records the agreement of terms and conditions at the beginning of a training post
4	**Assessments**	records which document progress and achievement throughout Foundation, as well as identifying problem areas so assistance can be provided
5	**PDP**	A record provided to structure and plan educational, and career, progress

Whilst trainees were encouraged to complete the Foundation Portfolio in its entirety, not all sections were mandatory: Assessments/Competencies, Educational Agreements and some forms were required by all.

Figure 9 shows a screenshot of one of the assessment forms, Significant Event Analysis.

Figure 9. Screenshot of Significant Event Analysis

4.1.1 Foundation Medicine in Scotland

NHS Education for Scotland (NES) is a special health board "responsible for the development and delivery of education and training for all NHS Scotland staff and for

supporting NHS services to the people of Scotland". Within NES, the commissioning and delivery of postgraduate medical education falls under the medical directorate, but NES also plays a significant role in the education and quality assurance of undergraduate medical education within Scotland.

Foundation Schools combine deaneries (who hold the responsibility for the delivery of postgraduate medical training, as well as continuing professional development, for all doctors and dentists) with trusts (outwith Scotland), health boards (Scotland), hospitals and other organisations to provide a wide range of training in varied settings. Deaneries have the further responsibility of providing and training educational supervisors for the Foundation years. Scotland constitutes a single Foundation School (one of 25 in the UK, 2011) but is comprised of four medical deaneries (North, East, South East and West). The four Scottish postgraduate deaneries have operational responsibility for ensuring that all aspects of postgraduate medical education, from Foundation to Core and Speciality training, are delivered to the highest standards.

The Scottish Foundation School provides a wide range of programmes delivered by the four deaneries offering a range of training experiences covering different types of populations (from teaching hospitals to remote and rural hospitals), numerous specialties, and geographically diverse areas.

The Scottish Foundation School was established in 2005 to deliver a taught programme to ensure trainees can meet the requirements of the curriculum. The variety of potential learning outcomes mandated a much more systematic approach to delivery and recording than employed in previous programmes. This included the development of a formal programme of education mapped against curricula content, in addition to appropriate induction and mandatory training tailored to requirements at each locality. All trainees were given access to a named educational supervisor and used the ePortfolio and e-learning systems to support their training and document evidence of their progress. Educational supervisors also used the ePortfolio to identify poorly performing trainees, and offer the appropriate assistance.

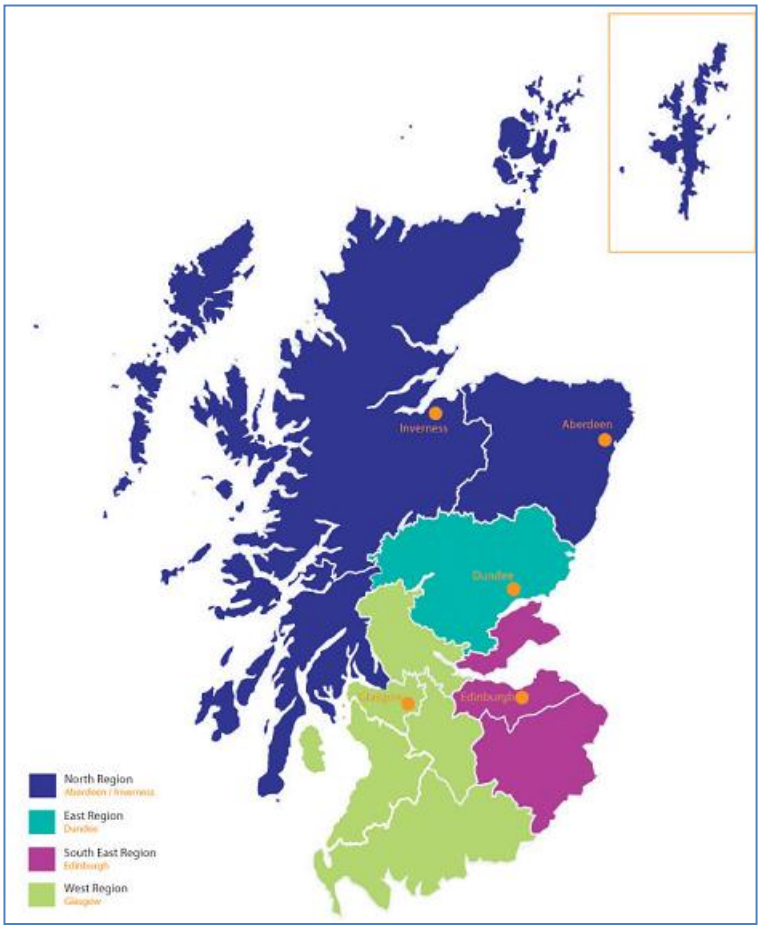

Figure 10. Map of Scotland Showing Medical Deaneries

Whilst the Foundation Programme is UK-wide and all trainees use the same curriculum, there are considerable challenges to delivering this programme to Scottish Foundation doctors due to the geographically diverse nature of the country (Figure 10). Over half the trainees are based in teaching hospitals (Aberdeen, Dundee, Edinburgh and Glasgow), but a significant number of Foundation doctors work in small remote and rural hospitals often with only a few other trainees (core medical training and general practice) and consultants and no senior trainees. This poses problems in terms of delivering the curriculum across the different sites in a consistent manner but also means that the respective Foundation Programme Director for a trainee can be in a completely different location. ePortfolio therefore provided essential flexible support to users and to Deaneries meeting their obligations to the regulator by providing consistent content and tools and enabling the delivery of uniform educational

processes.

4.2 ASSESSMENTS/COMPONENTS WITHIN EPORTFOLIO

There were three main types of assessment for UK Foundation in 2007-08, but there was significant variation within the United Kingdom with regards to the individual assessments used within each broad type and the frequency with which they were conducted. The ePortfolio recognised users' locations upon login and assigned the appropriate versions of the assessments according to their UK location. Within Scotland all four deaneries shared the same assessment tools and schedule.

To be recognised for practice, ePortfolio had to become the electronic equivalent of the paper copy provided as the "Foundation Learning Portfolio" by the UKFPO. The paper copy stopped being printed in 2008, but continued to be updated annually and available as a printable document (booklet) as a downloadable PDF until 2010. All assessments (Table 7) in Scottish Foundation were recorded in the NES ePortfolio, as well as the Personal Development Plan (PDP). The ePortfolio therefore also contained the educational agreement, statements of health and probity, records of meetings of supervision and career planning.

Several components of the ePortfolio were used to group the population and enable the evaluation of the core research questions as shown in 8.

Table 7. ePortfolio Components and their Evidence

ePortfolio Component	How it is Used
MSF (see Appendix 1)	Determining the population groups by self-assessment MSF
Educational Log / Significant Event Analysis	improving the accuracy of learner perception of their learning needs
PDP	promote a change in learner activity
Supervisor's Report	improve clinical practice

4.2.1 Multi-Source Feedback

Multi-source feedback (MSF) is also known as multi-rater assessment and 360-degree feedback. It is feedback that can come from any prescribed person (the UKFPO identified professionals and seniority) that can judge individual performance. In Foundation medicine this could be a variety of clinical and non-clinical roles, including supervisors as well as senior colleagues, senior nurses or pharmacists. Results were compiled, anonymised and returned to trainees for discussion with their supervisor.

Scottish Foundation programme leads developed their own MSF tool and as with many MSF systems, it included self-assessment with externally rated assessment. This was critical for this study as the self and peer assessment data could be compared from a common tool. Outside Scotland the UK used one of two similar tools, TAB (Team Assessment of Behaviour) or mini-PAT (Mini Peer Assessment Tool). As of 2010, all regions of the UK came to use TAB as their MSF tool.

4.2.2 Educational Log / Significant Event Analysis

The recording of noteworthy educational experiences was recognised as an important aspect of the UKFPO portfolio. Both the UKFPO paper and electronic e-portfolio versions contained an educational log to describe and record educational events in a structured format. The entries could be kept private to the trainee or shared with their supervisor. The educational log was designed to enable discussion of clinical reasoning, personal reflection and decision making in a supportive environment.

Significant Event Analysis (SEA) was established as an assessment instrument for general practice medicine, but has been adapted for more widespread use within medicine as well as other professions. In Scottish Foundation, SEA was both a type of event that could be recorded voluntarily in the Educational Log over and above the mandatory requirement for one peer reviewed SEA. SEA was designed so practitioners could learn from both the strengths and weaknesses of the care they provided. In SEA the individual identified an event of note to them and the tool gave a structured approach to analyse, discuss and reflect upon it. Additionally, SEA was designed to

identify best practice and facilitate communication between the individual and larger team.

During the selected training year for study (2007-08) the rest of the UK used Case-based Discussion in place of SEA, and this tool replaced SEA in Scotland the following year. Although the format has continued to change, all the UK continues to use a (now standardised) Educational Log.

4.2.3 Personal Development Plan

The PDP was the tool designed to help the trainee describe what they aimed to achieve in particular placements or over the training year. Though it was not mandatory, it was intended to be completed and discussed with the educational supervisor by the end of each post to ensure that the goals were being met. Guidance provided by the UKFPO at the time stated that the assessment tools would help trainees identify areas that needed work.

4.2.4 Work-Place Based Assessment

Postgraduate Medical Training Board (PMETB) (2007) described work-place based assessment as the evaluation of what a doctor actually did in the workplace. Generally it was conducted in the workplace itself. It could be initiated by both trainee or jointly by the trainer/assessor (the latter of which could hold a variety of different roles – listed in Table 9). Workplace based assessments aimed to evaluate the top two levels of hierarchy of the medical education assessment pyramid, i.e. “performance” and “action” (Miller, 1990), see Figure 11.

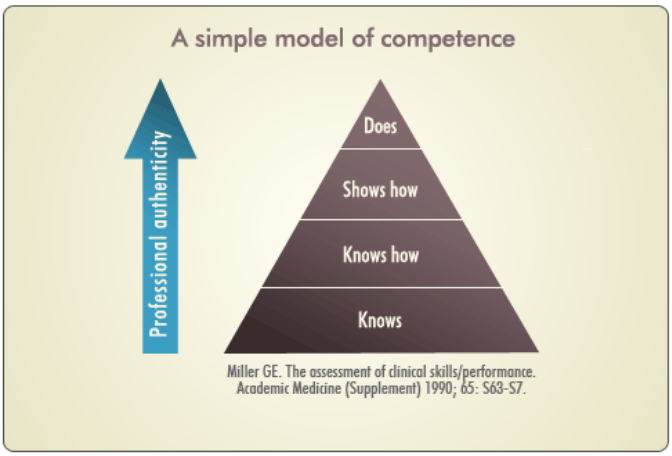

Figure 11. Miller's Model of Competence

The Scottish Foundation School opted for a generic work-place based assessment tool (simply called Work Place Based Assessment or WBA) to measure a trainee's progress over time in specified areas, both clinical and non-clinical. It provided a structured format and immediate feedback was given to the trainee on clinical encounters; these recorded scores for pre-defined skills on a seven point Likert scale (1=highly unsatisfactory to 7=highly satisfactory) by a range of approved individuals (consultants, senior nurses, pharmacists, etc.).

The assessment tools in use in the rest of the UK (Direct Observation of Procedural Skills/DOPS and Mini Clinical Evaluation Exercise/Mini-CEX) have come to replace the Scottish Foundation Work-place based assessments, but the approach and content remains broadly the same.

Overall, the great majority of trainees completed all fifteen and six clinical assessments required of them in each year, respectively. Among trainees who failed to submit all those required, the median missing number was one for first year and three for second year trainees. The quality of scores was generally very high with a median of 7 in three quarters of first year assessments and two thirds of second year, the rest were 6. Submitted assessments scored lower than 5 accounted for 1% of first year and 3% of second year trainees. These were not included in this study as they could not be used to answer the review's questions. It should be noted that the two years were not compared in this study, and arguably could not be as the content of each training year

is intended to be distinct in focus and application.

4.2.5 Required Content

The assessments described above form part of the wider minimum requirements for clinical and non-clinical activity to evidence satisfactory completion of the first and second years of the Foundation programme in 2007-08. The full details are shown in Table 8, along with the PDP and the two formal reports for competence – the Supervisor's Report and COP.

Table 8. Details of Foundation 2007-8 ePortfolio Components, Purpose, Frequency and Requirements

Assessment / Record Type	Question From Self-Assessment Review	Eportfolio Component Described In Table 7	Content / Purpose	FY1 Minimum Requirement	FY2 Minimum Requirement
Multi-Source Feedback	(identifying self-assessment quartiles)	4/Forms	structured assessment of 22 elements of professional & clinical skills by trainee-selected peers and self	4 peer + 1 self during post 1 and 3	4 peer + 1 self during post 2
Educational Log	perception of learner needs	2/Forms	self-directed semi-structured record of learning events (e.g. lectures attended, procedures conducted)	evidence of use throughout year	evidence of use throughout year
Significant Event Analysis	perception of learner needs	4/Forms	type of structured record in the Educational Log - trainee-selected incident used to promote reflection and evidence of implementation of learning	1 shared & reviewed by Educational Supervisor during post 2	1 shared & reviewed by Educational Supervisor during posts 1 and 3
Personal Development Plan (PDP)	change in learner activity	5/PDP	self-directed semi-structured record of plans for personal development and actions taken	evidence of use throughout year	evidence of use throughout year
Supervisors Report & Certificate of Performance	Improve clinical practice	2/Forms	formal structured record that appropriate level of competence was achieved during post	1 each per post	1 each per post
Workplace Assessments	Improve clinical practice / patient outcomes (not answered)	4/Forms	21 defined clinical assessments (e.g. FY1: initiate IV infusion, FY2: advanced life support)	n=15	n=6

4.3 NES ePortfolio

In advance of the 2005 introduction of the Foundation Programme in Scotland, NHS Education for Scotland (NES) designed and implemented a pilot web-based electronic portfolio for 400 first year trainees in the South East and North deaneries. (A smaller simultaneous pilot evaluated an e-portfolio on the same platform for 44 GPST trainees.)

> The role of the author was initiating, planning and managing the pilot project in partnership with the Foundation Manager of the South East deanery (Edinburgh).

The pilot was to determine whether an electronic portfolio was viable and offered any advantages over the paper copy. It was developed with the intention of allowing structured recording of training activity, facilitating interaction and flexible access to robust information for educational supervisors, programme directors and administrators. The system's database was in SQL (standardised query language) with a simple web interface designed to present and manage trainees' evidence throughout each post of their training programme (and beyond). As described in Table 8,section a number of mandatory assessment records would be collated via ePortfolio providing a cumulative record of evidence of self, peer, and supervisor assessment of their competence during the year and allowing regular review by their supervisor.

The pilot was internally evaluated during the first six months, using a survey, interviews and focus groups, as well as an analysis of usage data. This was compared to the paper portfolio system that was being run concurrently. The in-house evaluation (unpublished paper, correspondence) suggested the electronic version demonstrated efficiency savings, enabled superior quality assurance processes and saw higher completion rates than the paper version. There was strong growing demand from trainees using the paper version to be allowed access to the electronic system. This

was particularly notable in boundary areas between deaneries (e.g. hospitals that could have trainees from both paper and electronic portfolio deaneries). On this basis due to the positive initial feedback and the building demand amongst trainees, the NES ePortfolio became a permanent part of Scottish Foundation and from August 2006 was extended to support all first and second year trainees (n=1600) in all four deaneries (Table 10).

Also in August 2006, the Foundation Medical ePortfolio expanded to include Wales, Northern Ireland and several English deaneries. Adding each of these areas required customisation, as did the implementation of different local assessment tools, forms and processes. Additionally, in conjunction with the Joint Royal Colleges of Physicians Training Board (JRCPTB), the first pilot was launched for higher specialty training in Merseyside. The NES ePortfolio also spread within Scotland with two versions for Pharmacy and a sophisticated build for Dentistry with the electronic version of the RPA (Record of Practice and Achievement – required for satisfactory completion within dental vocational training).

The ePortfolio continued to expand in terms of both numbers and types of users. The broader range of professional groups using the system had numerous advantages, such as the sharing of good practice. A single common system also allowed users to have different roles within a system; for example, a pharmacy educational supervisor might also be asked to conduct a multi-source feedback on a Foundation medicine trainee. A common codebase allowed for shared development as well as practice, with groups learning from others' experiences in monitoring and quality assuring trainees for their own purposes or for regulatory obligations.

4.3.1 ePortfolio Technical Summary

This small section describes the technical detail (structure and functions) of the NES ePortfolio during the selected training year selected for study (2007-08). This was the first version of the software and was in fact, coincidentally, an expanded and heavily modified adaptation of the software used by BEME reviewers, including the those described in Chapters 2 and 3. The original BEME software allowed reviewers to code,

collate and agree scores for their papers on a web based form-driven system, and access all data, as well as user activity through reporting functionality. This original basic functionality was extended and adapted in creating the original ePortfolio for Foundation Medicine.

4.3.1.1 Architecture

Two applications of Foundation Medicine ePortfolio existed in the 2007-08 training: Scotland and the rest of the UK. The Scottish version was linked (via database table unions) to the DOTS (Doctors Online Training) e-induction/learning system, synchronising user details and trainee posts, and was used only by Scottish foundation users. The UK ePortfolio was used by non-Scottish Foundation medical programmes, all (Scottish/UK) Core Medical Training (CMT / Royal College(s) of Physicians), Royal College of Paediatrics and Child Health, Scottish Pharmacy, and Dentistry. The UK version, though derived from the original Scottish version, was hosted on a different web server with a distinct separate code base, and drew data from a distinct database, which simplified the extraction of data for this study.

4.3.1.2 Technology

The initial version of ePortfolio used well-established technologies, as time and budgets did not allow for technical innovation. ePortfolio v1 was coded in VBScript ASP 3 (“Classic ASP”) accessing data from a SQL Server 2000 database on a Windows 2000 server platform. The database was a single tier set up, i.e. data requests were made from the presentation layer pages rather than from a business logic or data access layer. When the application was re-written in 2008 there was the opportunity to embrace cutting edge technologies to better prepare for future development.

4.3.1.3 Key Features

The ePortfolio was hard-coded (written directly into the core system and therefore very difficult to modify) to support the basic role types: Deanery Admin, Regional

Admin, Educational Supervisor, Tutor (Programme Director) and Trainee. When an individual logged into the ePortfolio their unique identification authenticated them to the system and their role or roles were assigned; for example, one could be an Educational Supervisor and a Tutor, or have identical roles within other version of the ePortfolio (Foundation, Physicians, Paediatrics, etc.), see details in Table 9.

Table 9. Description of ePortfolio Roles

Role	Access rights
Deanery Admin	Ability to administer trainee accounts across Deanery; access to trainee ePortfolios within area
Regional Admin	Ability to administer trainee accounts within Trusts or Hospitals; access to trainee ePortfolios within area
Educational Supervisor	Access to prescribed areas of their trainees' ePortfolios (not private areas)
Tutor (Programme Director)	As educational supervisor, but wider population of trainees
Trainee	Full access to individual ePortfolio

Locations for each role were limited to three layers of description as shown in Figure 12.

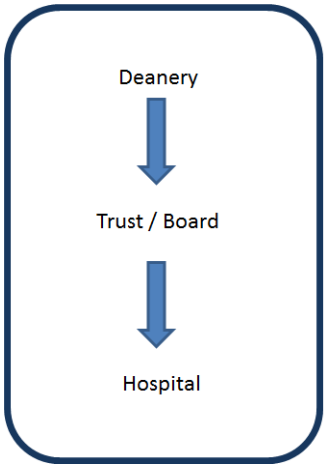

Figure 12. Role Hierarchy: ePortfolio v.1 2005-08.

Within the trainee role, posts were allocated to an individual. These were based on sequential dates as a single layer i.e. there was no use of a "parent" post to describe a

training period. One supervisor was directly associated with each individual post (only the Royal College of Paediatrics and Child Health (RCPCH) version's code allowed more than one supervisor per post). The post location was used to define trainee lists for administrators. Designated administrators also had the ability to generate new users and posts, as well as editing them.

All forms (assessments, declarations, reports) were generated for display from a database-derived set of individual form control elements (e.g. text box, radio button). Form data was saved to either individual form-specific database tables or were saved to a generic form data table.

Three distinct processes were used to display curricula for Foundation, Core Medical Training (Physician), and RCPCH, and each used differing methods to record comments, ratings, and associated files/forms. The Foundation ePortfolio in 2007-08 had a static display-only curriculum. Less cumbersome and more efficient curriculum functionality came with the introduction of ePortfolio version 2 in August 2008.

Users were authenticated to the system using a standard unique identifier and password system. The ePortfolio had separate internal messaging and support systems. Basic standard status reports (to give snapshots of activity, completion rates, etc.) were designed in advance and allowed administrators (or system administrators) to identify and interrogate relevant data within the system's hundreds of tables.

4.3.1.4 Growth of System After 2006

The initial version of the ePortfolio was not designed to be a portfolio system, but instead was created as a data abstraction (and, to a lesser extent, synthesis) tool for the BEME Collaboration. However, as the initial Foundation medical pilot team had very little time or resource to create or procure a new purpose-built system, this research software was adapted to become the version one ePortfolio.

The initial system shared two critical components with an e-portfolio system: it was form-driven and it allowed different roles (lead researcher, reviewer, admin) to be created, each of which could be assigned different access rights. This base system was fairly readily adapted to become the ePortfolio, and it rapidly evolved through regular

change requests upon going live. However, in its rapid transformation from in house research interface to e-portfolio, all emphasis was placed on ensuring users could readily and securely store data; retrieving that data for analysis was not considered for inclusion in development until the design of the second version of ePortfolio, a fact that would have a large impact on the potential to analyse these data.

Brisk expansion of the ePortfolio continued in 2007, and it became apparent that the system, which was designed for a small pilot, was simply not best suited for long term development. The rapidly rising numbers of users and their demands for increased functionality contributed heavily to this. But an equally strong driver was the fact that the development time involved in adding blocks of code to a system initially designed for less than a dozen BEME systematic reviewers was untenable, so a new platform was created. The new ePortfolio (version 2.o) would also give the team the opportunity to draw upon the experience gained over the first years to design a system specifically intended to be an e-portfolio, rather than continually modifying something that was not.

Each year since then, the ePortfolio has seen consistent growth in the number of organisations/groups, and corresponding users, within the system – the annual growth is summarised in Table 10.

Table 10. Chronological Growth of NES ePortfolio

Training year	User groups added per year	Number of users*
2005-06	Foundation One pilot, two Scottish deaneries	710
2006-07	All Scottish Foundation as well as Wales, Northern Ireland & 3 English deaneries; JRCPTB pilot (Mersey); Dentistry (Scotland); Pharmacy (Scotland)	7,400
2007-08	Further English Foundation deaneries; JRCPTB (UK roll-out); Royal College of Paediatrics and Child Health; Scottish Dentistry expanded	45,000
2008-09	Royal College of Obstetrics and Gynaecology; College of Emergency Medicine	57,000
2009-10	Remainder of English Foundation; Royal College of Radiologists; Faculty of Sexual and Reproductive Healthcare; Government of Malta; Scottish Nursing and Midwifery pilot	90,000**
2010-11	Roll out of Scottish Nurse Mentor	138,000
2011-12	Royal College of Physicians of Ireland (training and revalidation, two versions); Medical undergraduate pilot (UCL, Glasgow, Bristol, Brighton & Sussex)	180,000
2012-13	Scottish Dentistry join main system; expansion of undergrad (more years added, Keele, Queen's, Cardiff join); Faculty of Occupational Medicine; Faculty of Intensive Care Medicine; Malta General Practice; Royal Navy; external assessors register many more accounts.	245,000

** Approximate. Includes all systems roles (trainee, supervisors, admin, etc.)*
*** From 2009 external assessors were able to create their own accounts*

For August 2008 the entire system was rewritten in updated technologies (.NET 3.5 and SQL 2008) that became ePortfolio version 2.0, to serve the expansion and diversification of NES ePortfolio, but also to facilitate the increasing demands for functionality that came from the users. These included a dynamic curriculum, customisable reporting, file upload and the ability to link and collate the different items (assessments, reflections, guidance, etc.) within an ePortfolio.

Version 1 (v.1) of NES ePortfolio allowed basic pre-designed queries to interrogate the data, but version 2 (v.2) was designed so any permitted user could view and analyse particular sections could readily. ePortfolio v.2 was designed so data was not only easily accessible but also easily exportable whilst still retaining its security to unauthorised access.

Whilst the Scottish Foundation training year (2007-08) provided a dataset of enormous

depth and detail, the ePortfolio was primarily designed to store data securely, rather than for analysis, which resulted in further challenges to address the questions. Therefore, this research required considerable effort to retrieve, cleanse and verify data from a system that and experienced massive unexpected growth for three years.

5 CASE STUDY

This case study enabled the exploration of the self-assessment review's questions on a large data set of an entire training year, with three of the four self-assessment review research questions matching ePortfolio component areas:

I. **Improve the accuracy of learner perception of their learning needs?**
II. **Promote an appropriate change in leaner activity?**
III. **Improve clinical practice?**

As there had been no published analysis with such a dataset, initially this study explored the extent to which quantitative and qualitative analysis could be employed.

Once the data were extracted, it became apparent that the way the ePortfolio was used in this training year meant that purely quantitative methods were unlikely to adequately interrogate the data and statistical analysis of the data was unlikely to be meaningful. While the quantitative data, such as MSF scores, allowed allocation of trainees to quartiles and comparison to peers, it revealed little about trainees' development and competency. Comments, while more time consuming to analyse, added richness to understanding the individuals and the groups in which they were classified. Comments were reviewed in full, and key words identified (for example positive and negative terminology, depictions of action or reflection on behaviour), these were themed and grouped and their frequency within each self-assessment group was assessed against the published literature.

The self MSF provided scores and dates which could categorise the trainees by assessments scores into the Kruger and Dunning quartiles early in the training year.

The self-assessment review also identified a number a number of areas where it was felt there was considerable evidence, which could now be tested with ePortfolio data. These areas included several that fell within the Kruger and Dunning's work (insight, improving self-assessment, comparisons with peer and faculty, and novice versus expert). But it also included issues beyond their work, such as differences in self-assessing technical and "soft" skills, the role of feedback and summative versus formative applications.

Other areas had less scope for investigation due to the data available. There was inconclusive and/or contradictory evidence in a few areas including: gender, culture, and the purpose of task self-assessed, but no data readily available in the system to investigate these factors. Other issues could not be readily tested without considerable further work, including the use of video feedback and benchmarking, a formal evaluation in the applications of instruction in both self-assessment as well as the particular skill, and the concept of the experience (or even skill) of self-assessment. It was also beyond the scope and ability of this work to formally examine the acceptability of self-assessment itself, though this will be discussed in the final chapter. The portfolio review's findings employed to inform the case review's methods as described in Chapter 3. Furthermore, the wider review findings inform about the extent to which an e-portfolio can be used to facilitate self-assessment, and these issues are discussed in the final chapter.

Both the self-assessment and portfolio systematic reviews were challenged by having an evidence base that mainly consisted of studies with varying populations, quality and interventions and/or were so small in scale that generalisability to the wider world was limited.

This chapter draws upon a primary set of real life data of a considerably larger scale than previous published studies, namely the UK Foundation medicine programme's ePortfolio.

The ePortfolio platform provided a natural laboratory to revisit the reviews' questions via the data collected by an enhanced electronic platform which recorded assessments, self and external, in the training year (August 2007-August 2008) selected for study. Evaluation of self-assessment in relation to assessments by others/non-self was possible, as well as in relation to other educational events and opportunities. Additionally, the medium of e-portfolio could be judged as a tool for recording assessment across a twelve month training period. The data recorded about these trainees was extensive not only in the numbers (1600 trainees) but also its depth – the system logged every entry's activity and duration, as well as the interaction (e.g. messaging, sharing of records, etc.) with other uses, including supervision. Though

unwieldy, the data provided an enormous amount of detail that was unavailable in the published self-assessment or portfolio literature, as described in Chapters 2 and 3.

5.1 DATASET

The self-assessment systematic review identified the need for complete, robust and comprehensive datasets to provide an adequate evidence base for examining the subject area. From the ePortfolio, a dataset from an entire training year offered the opportunity to attempt to address that review's central research questions about the effectiveness of self-assessment in improving learner perceptions of their learning needs, promoting an appropriate change in learner activity, improving clinical practice and improving patient outcomes.

The case study involved a large scale retrospective examination of self-assessment data gathered for purposes other than the study of self-assessment itself, namely the monitoring of progression of trainee doctors and their personal and professional development.

5.1.1 Data Extraction

As each ePortfolio record was associated with a single trainee in a specific post, the dataset provided a national overview of the range of posts included in each year of the Foundation programme, including details such as speciality, region and duration.

From the full ePortfolio database containing all submitted records for all users since 2005, the population of interest and their relevant records were identified by a number of definitions:

- designated role; foundation trainee, excluding specialty trainees, supervisors and admin etc.,
- Scotland based posts; excluding users in England, Wales and Northern Ireland,
- training year 2007-08; excluding records submitted by the above trainees at different times, and records for users who were foundation trainees at different times.

The ePortfolio supported trainees over extended periods in which movement between posts and geographic locations happened several times every year. Within each placement a range of educational experiences were recorded, which could have regional variation in requirements, as well as start and end times. The queries interrogating the database therefore had to select nominated time periods in which relevant roles were held to include the submissions intended for analysis (e.g. the duration of the first post of Foundation One).

5.1.2 Data Cleaning

Once the study population was identified, all records (completed or partial) submitted for or by those individuals within the study time period were extracted using SQL queries and exported for processing into Excel as individual spread sheets for each item type (e.g. educational log, workplace based assessment). Further analysis was undertaken in SPSS (PASW18).

In preparing MSF data, it was necessary to add a code to identify self and non-self records as this was not inherent in the form at the time (it has since been added). This was done in two ways to try and capture all self-assessments accurately: by comparing the GMC number of the assessor and assessee (where provided); and secondly if the word 'self' appeared in the role of assessor. It is possible that some records were not accurately completed and so have been miscoded. A review of numbers indicated that this identified the expected volume of MSFs, as defined in the Foundation requirements that specified required submissions, taking into account the expected regional variations (e.g. in one region additional forms were requested of trainees).

In place of the individual's GMC number an anonymised unique identifier was used to maintain confidentiality during analysis. Data were stored within a password protected NHS network, and on encrypted IT equipment when off-network.

This work developed from a study of the process details of ePortfolio use, for which an ethics adviser noted that as it comprised analysis of routinely collected information, its evaluation did not require ethical approval. Guidance was issued that Research and Development leads in each Health Board should be informed about the ongoing use of the data for research and evaluation purposes. This was duly done.

As data were extracted from the complete working database (which retains everything ever entered) there were a number of processes required to clean and prepare it for analysis. This included, for example, removing “process” scores within the Multi-Source Feedback which did not indicate an actual assessment value: zero, which indicated an incomplete record and eight, which indicated that the assessor had not observed the trainee sufficiently to enter a score. These scores are informative, but do not offer a useful ordinal value according to the one to seven Likert scale, for valid completed assessments.

The retention of incomplete records, while useful for trainees, was a complication which required attention, as did the presence of duplicate records (required due to frequent inadvertent or accidental resubmission of the same record with multiple clicks of the submit button). These were checked by sorting records and using comments fields or the time stamp of records to identify unnecessary repetition. Where possible such records were removed to ensure a fair interpretation of the volume and content of data submitted for each individual. It was possible however that some duplication persisted in aggregate analysis; however, at the individual level such duplication was readily spotted.

In a small number of cases, assessors misinterpreted the order of the scoring scale, and scored trainees low instead of high. Where comments were provided this was immediately obvious, and in these cases these were transposed to the top of the scale as appropriate. It is possible, however unlikely, that some (without comments) were missed, as very low scores were always supposed to be submitted with comments, therefore these would be readily identifiable.

5.1.3 ePortfolio Usage and Implications for Analysis

As will be described later, analysis of assessment data contained in the ePortfolio indicated that trainees and supervisors predominantly used it to demonstrate the ultimate competence required to be achieved during their training programme, rather than to track their progress towards competence. This is a legitimate purpose and meets the mandated *final* requirements set out by the Foundation programme but it does differ from some other e-portfolios and the overall guidance of Foundation which encourage users to record assessment data throughout their learning process documenting improvement from below, up to the expected standard over time.

This has implications for analysis, as the assessment data contained in this ePortfolio is not normally distributed. MSF scores, for example, tended to be skewed heavily towards high values, with the vast majority of trainees giving themselves sevens and sixes. This meant that it was not possible to sensitively identify groups of high or low self-assessors from individual reports.

As this was a real dataset, in that trainees and supervisors entered data in real time, there was variation in adherence to data submission requirements, therefore the data submitted for each trainee is not uniform and complete. However, an audit of the completion of the data submitted has been undertaken (Tochel, Beggs, Haig *et al. 2011)*. The ePortfolio was specified to allow this in these early days where the Foundation programme was being implemented across diverse local environments of the four deaneries (at this point, time periods were not locked down/or rigidly adhered to). As time progressed in the use and development of ePortfolio, submission dates and other entries have become progressively more fixed. Variations in submissions were relatively common in all system roles, with submission dates often missed by days or more. The number of actual submissions for required forms also varied; many of these variations could possibly be put down to the progress different geographical areas were making in adhering to the new Foundation standards, but analysis of this was out of the scope of this thesis. There was also variation in the degree of competence of users with the new assessment tools and scoring systems, as well as (for some) the experience of using an online system itself. Therefore, it was not

uncommon to discover erroneous or incomplete data (as already mentioned), and where possible corrections were made or validation sought.

Finally, there was the potential for trainees to fraudulently enter their own scores. Although the system had built-in safe-guards and deaneries conducted random spot-checks on assessments, a compromise was required between maximum system security and usability. This was reached in consultation with stakeholders, and continues to be monitored online with new IT standards and ever changing user demands.

5.2 Analysis of Foundation Data

Initially, self-assessed MSF scores were used to identify the key populations as defined by the self-assessment literature: the high and low self-assessors amongst the four quartiles that demonstrated the recognisable characteristics of over and under self-estimation.

To tease out groups of high or low self-assessors therefore, all clinical self-MSF scores were collated early and late in the year (as self-confidence or efficacy may change over time). As these scores constituted categorical data (i.e. whole values between one and seven) and as mentioned in Section 5.1.3 the majority of scores were at six and seven, this did not allow a clear distinction of trainees into quartiles. Therefore an early mean value was calculated for each individual (see Table 12), thereby creating a continuous variable which could be used to rank trainees, weighted to reflect the frequency and consistency with which they had submitted scores at a given level. More weight was given to trainees who had repeated assessments – i.e. consistent high or low scoring was ranked higher than one-off high or low scores.

Clinical self-assessment scores were extracted and ranked. The high and low self-assessment groups' scores were compared against their supervisor's ratings to see if they corresponded.

Next the ***first*** of the review's questions asked if self-assessment "improved the accuracy of learner perception of their learning needs"– the questions were addressed by each of the three core assessments in the ePortfolio, as well as other specific

evidence collated by the system. These logs recorded any chosen educational event, and could be kept private or shared with their supervisor. The logs prompted trainees to reflect upon these events and determine their learning needs. This analysis looked to see if self-assessments did indeed go onto influence the accurate evaluation of individual learning needs.

Personal Development Plans (PDP) were examined to see if self-assessment could "promote a change in learner activity", the review's ***second*** question. PDP forms allowed planning of individual's learning in a structured format and recorded when plans were achieved.

The ***third*** question looked at supervisor reports as the ePortfolio's closest measures of "improved clinical practice". Self-assessments were evaluated in relation to these, to determine whether there was any correlation to completion of the forms.

No attempt was made to answer the review's fourth question (improved patient outcomes), as the Caldicott principles preclude the recording of any patient identifiable information in ePortfolio, therefore no evidence was available.

5.2.1 Defining Self-Assessment Groups for Comparison

Since Kruger and Dunning (1999) first described self-assessors falling into distinct predictable quartiles (detailed in Section 2.2), numerous other studies have found the same, in healthcare as well as unrelated settings. The ePortfolio dataset provided a natural laboratory to test whether this would be replicated across a large group of junior doctors. But additionally it afforded the chance to determine whether clinical self-assessments were more accurately conducted than non-clinical, as the MSF tool measured both sets of skills.

In 2007-08 self-assessment MSFs were required in each post 1 and 3 (first year) and 5 (second year) of Foundation training, though trainees could submit additional ones at any time. The requirement was met by 91% (1st year) and 90% (2nd year) for self MSFs and 85% (1st year) and 82% (2nd year) non-self. Four non-self MSFs are also gathered in posts 1, 3 and 5. As the first and third post mark the beginning and end of the first structured training year, these periods were chosen to examine the potential changes

in scoring. Trainees (n=1604) were required to submit a minimum of four non-self and one self MSF during two first year posts and one second year post.

Each MSF recorded the rater's score for 22 areas of the trainee's professional competence and clinical skill category and one global rating. Raters had the option to indicate "not applicable" if they felt they did not have the opportunity to observe the particular skill(s). Each category could be scored between 1 (highly unsatisfactory) and 7 (highly satisfactory). The multi-source feedback tool comprised twenty three statements about the trainee that were completed by both self and non-self raters, the final one being global.

The self-assessment systematic review revealed there was some evidence that clinical skills are more accurately self-assessed. It was thought to be because they are more tangible and less open to subjective interpretation. For the purpose of this study, within the Scottish Foundation MSF, statements were reviewed to identify those which could be classified as "clinical" behaviour and practice.

The breakdown of clinical and non-clinical MSF statements was interpreted as follows: the clinical questions had the trainee participating in the direct care of patients, with impact on clinical outcomes, which could be objectively measured, (see

Table 11). From the segregation of the questions a detailed analysis of the data became possible.

Table 11. Components of MSF Assigned as Clinical or Non-Clinical

Category	Measure
non-clinical	The doctor is polite to patients The doctor is caring of patients The doctor is respectful to patients The doctor shows no prejudice in the care of patients The doctor communicates effectively with colleagues The doctor has a command of the English language at the appropriate level for patients Doctor respects others role in health care The doctor works constructively in the health care team The doctor is accessible to responsibilities The doctor demonstrates commitment to their work in the team The doctor demonstrates competence in problem solving The doctor constructs appropriate management plans The doctor seeks help where appropriate The doctor maintains an appropriate clinical record The doctor is professional in their work The doctor is in a state of health fit for practice
clinical	The doctor is routinely able to take a structured history from the patients (carers) The doctor is able to conduct examination of the patient in a structured manner taking full account of the patients dignity and autonomy The doctor is able to promptly assess the acutely ill or collapsed patient The doctor is able to appropriately manage and monitor the acutely ill or collapsed patient The doctor is able to prescribe safely and appropriate The doctor demonstrates competence in emergency care
global	The doctor's overall performance

5.2.2 Extracting Group Data

Data on each of the six self-assessed clinical skills identified in Table 11 were extracted from the full dataset and the distribution of scores examined. Scores were positively skewed, with most at six or seven. Therefore to sensitively identify quartiles of high and low self-assessors the mean of all six clinical scores for each trainee at the start and end of the year was calculated. This was done by grouping individual trainee's self MSFs submitted during each post. These mean clinical scores were more evenly distributed and therefore allowed assignation of trainees into categories defined as low (bottom 25%), mid (central 50%) and high (top 25%) self-assessors. These assigned self-assessor categories form the basis of the rest of this paper.

Trainees with sufficient self-assessment submissions to allow analysis over the academic year were identified. The requirement was one or more self-assessments in the first four-month submission period of the academic year (15/07/07 – 31/12/07), and one or more in the third four-month period (01/05/08 – 01/08/08) of the academic year. Although a small number of posts started and finished at unusual times, for the purpose of this analysis, these four-month submission periods were used to group early and late assessments, which in the majority of cases coincided closely with trainee posts.

In each submission period, the scores trainees gave themselves for six clinical categories of the MSF were extracted (see Table 12). As described in Section 5.2 this allowed the calculation of the individual's clinical mean self-assessment score at the start and end of their first training year, whether one or more forms were submitted. As previously mentioned, scores of 8 were not counted in the average score as they would distort it, but the number was counted, in case there was a relationship between trainees choosing not to score themselves, and their relative self-assessment levels. This differs from trainees whose submissions included zeroes - these were excluded as this indicated an incomplete submission. The analysis therefore focuses on the subset of trainees with one or more complete sets of clinical MSF self-assessment (and numbers therefore vary between each submission period).

As described in Section 5.2 the relative position of individuals ranked among their peers was noted in early and late self-assessments. The sensitivity of the ranking was enhanced by counting a combination of mean and number of submissions, i.e. if someone submitted 4 self MSFs and had a mean of 7.00 this is ranked higher than someone with one self MSF and a mean of 7.00 to reflect the consistency of the high mark. The true validity of this ranking as an indicator of trainees' self-assessment relative to their peers was not possible to verify objectively from the available information, but it forms the initial stage of exploration of this case study.

The tags "high" and "low" from this point will refer to trainees according to their initial self-assessments ranked among their peers.

Table 12. Groups Defined by Initial Self-Assessment Scores in Post 1

SA group	No of indiv-iduals	Submission period	No of valid scores	Mean clinical MSF scores			
				Group mean	std. dev.	min	max
all (Group A)	781	early period	730	5.7	0.6	3.3	7.0
		late period	739	6.1	0.6	1.2	7.0
one SA only	80	early period	35	5.7	0.6	4.7	7.0
		late period	44	6.0	0.5	5.0	7.0
high early (Group B)	162	early period	162	6.5	0.3	6.1	7.0
		late period	162	6.4	0.5	4.1	7.0
low early (Group C)	188	early period	188	5.0	0.3	3.3	5.2
		late period	188	5.8	0.5	4.2	7.0
mid early	345	early period	345	5.7	0.2	5.2	6.0
		late period	345	6.1	0.6	5.2	7.0

The minimum and maximum scores indicate the lowest and highest mean scores for an individual in that group. There were six trainees who only submitted self-MSFs between the early and late periods and therefore did not fall into any of the above groups.

The following section will describe peer MSF scores across the first year population and then in the subgroups as described above.

5.3 RESULTS

5.3.1 Data

A total of 14,878 MSF submissions were entered in the 2007-08 training year, of which 3,172 were self-assessments and 11,706 assessments scored by others. Both self and non-self clinical and non-clinical assessment scores were very high with medians of 6 or 7 in all of the 23 competencies.

A slightly higher proportion of first year (compared to second) trainees completed both self and non-self MSFs in 2007-08, though a sizeable minority did not; in consequent years requirements came to be met by nearly 100%. (See Section 7.3 for further discussion).

The range of mean scores in each type of MSF (self vs non-self, clinical vs non-clinical)

is shown in Figure 13 for the cohort of the first year. The global self-rating had a median of 6 and a mean of 6.12 whilst non-self had a median of 7 and mean of 6.51.

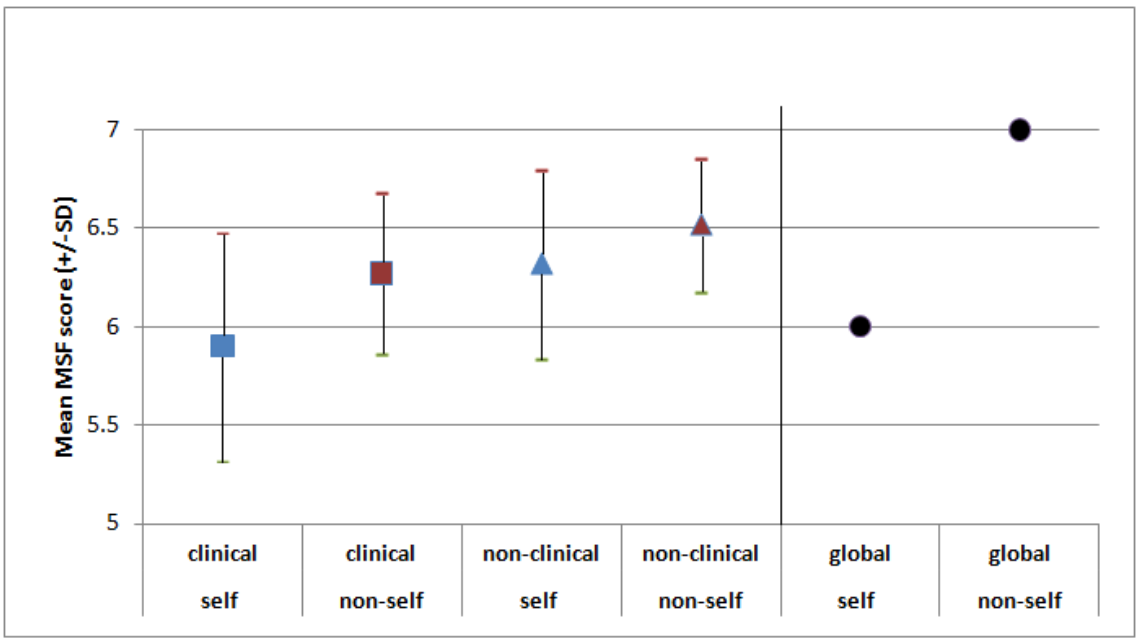

Figure 13. Mean and median MSF Scores for group A (all) by Subcategory of MSF

Whilst the raw self-MSF scores did not clearly delineate the quartiles as readily as described in other studies, the mean scores did so (t-test low vs high early, $p<0.0001$).

5.3.2 Sub-Groups

As previously mentioned, sub groups of the 2007-08 Foundation year cohort were identified as the "natural laboratory" test groups for comparison based on the published literature (the highest and lowest self-rating quartiles first identified by Kruger and Dunning). Their ePortfolio scores for assessments and other evidence will be compared in the following sections.

The total population (first year foundation trainees, Group A) will also be referred to for comparison where possible, for example trainees with one or more self-assessments in periods 1 and 3 who fell into the mid 50% (n~+400); and trainees with a missing self-assessment in period 1 or 3.

As part of the exploratory process, a number of subgroups were identified from the test group population where data may provide some more insight into the unanswered self-assessment review questions. All were explicitly defined from the population of trainees who submitted self-assessments in the first and last posts of first year. These were:

i. Group B and C: the high and low self-rating quartiles (defined by self-MSF) from the total for mean self-assessment MSF (test group, i.e. Group A)
ii. Group D and E: extreme self raters, i.e. group D comprises the highest scoring 10% from group B (n=30) and group E comprises the lowest scoring 10% scoring from group C(n=29)
iii. Group F and G: those within Group D and Group E who were rated contrarily by Supervisor's Report in Post 3 (i.e. a low self-rater who the supervisor scored highly; a high self-rater who the supervisor scored lowly)
iv. Group H: trainees who commented on their SA in both posts

The following sub groups are defined for further analysis.

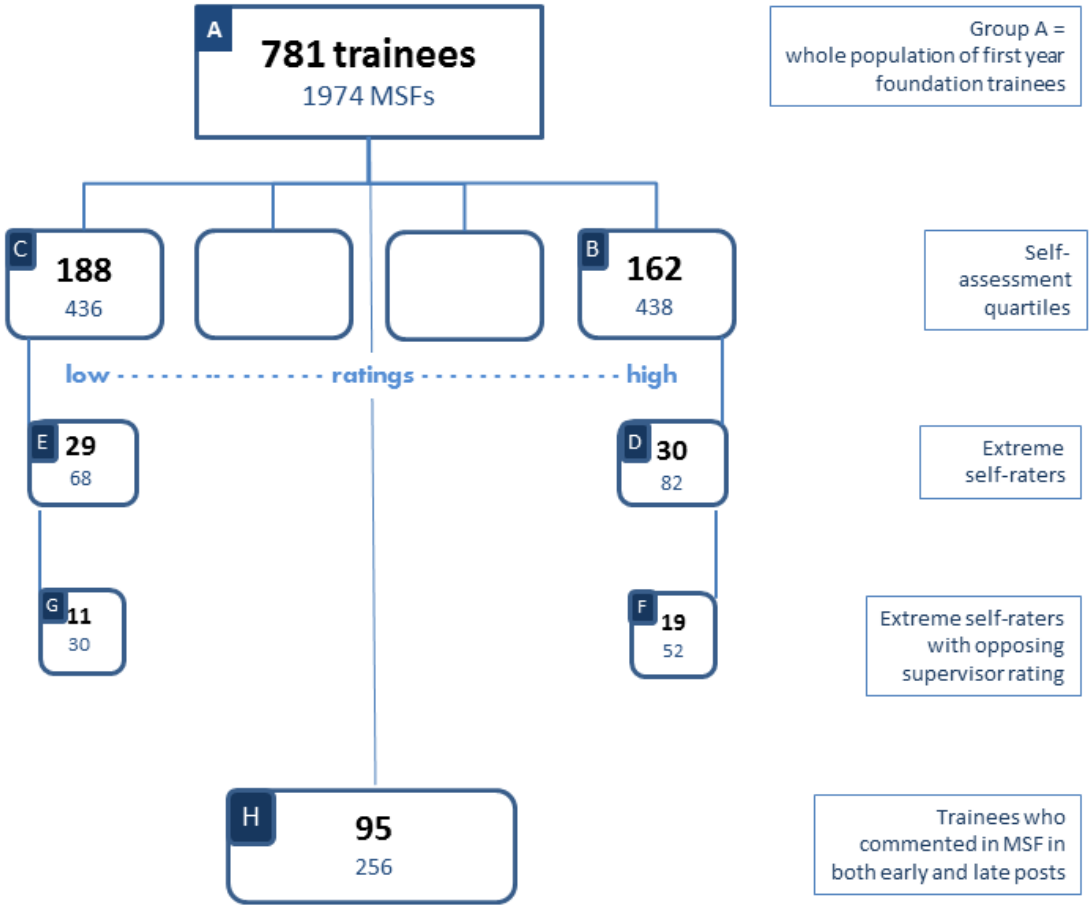

Figure 14. Sub-groups of Trainees by Self-Assessment Scores

Table 13 Percentage of Trainees Within Region by their Self Assessment Group

NHS region	one SA only	high early	low early	mid early	Total
East	1%	27%	21%	51%	**100%**
North	6%	13%	37%	44%	**100%**
South-East	12%	31%	13%	44%	**100%**
West	13%	18%	26%	43%	**100%**
NULL	0%	50%	0%	50%	**100%**
Total	**10%**	**21%**	**24%**	**45%**	**100%**

Table 13 reveals an example of the variation between Foundation regions/deaneries. This variation can be attributed to differences (setting of placement, geography, number of trainees, etc.) between each area, as well as how Foundation was implemented (it was not uniform in approach or comprehensiveness) in each, and is addressed in the Discussion. Records in which the region was not properly completed

were noted as 'NULL'.

5.3.3 Self-Assessment Status Change Between Posts 1 and 3

To examine how self-assessment changed within a training year, the scores from posts 1 and 3 (first and last post of the first year of Foundation) were compared. 775 trainees had at least one self-assessment submission in both these posts, totalling 1818 self-assessment records for the training year (a further 69 self-assessments were recorded in one post, but not the other).

The first three bars on Figure 15 show the self-assessment category (low, mid or high) into which early low self-assessors moved by the late period. Early mid and early high self-assessors are shown in an identical way to the right. As the figure shows, the majority of trainees remained in the same self-assessment category (relative to their peers) between post 1 and 3 (low/low n=99, mid/mid n=186, high/high n=80). 76 trainees moved from low to mid and 13 from low to high. From the mid category, 83 trainees fell to low in post 3 whilst a similar number (n=76) rose to high. 61 trainees dropped from high to mid with a much smaller number (n=21) falling to low. The number of trainees in each movement category therefore broke down quite predictably (in line with the self-assessment literature), with the single greatest number remaining in mid/mid between posts and the smallest numbers migrating between high and low (or vice versa). In order to validate the meaningfulness of these group movements and the relative positions of trainees to their peers, information from outwith the ePortfolio would be required, such as interviews with a representative sample of this population across the groups or information from supervisors with knowledge of trainees in multiple groups. Such research was beyond the scope of this study.

The mean clinical score below which low self-assessors fell, increased from 5.17 in post 1, to 5.83 in post 3 and 6.00 in post 5 (second year trainees) which may depict the recalibration effect described in the literature, and is covered in more detail in the Discussion.

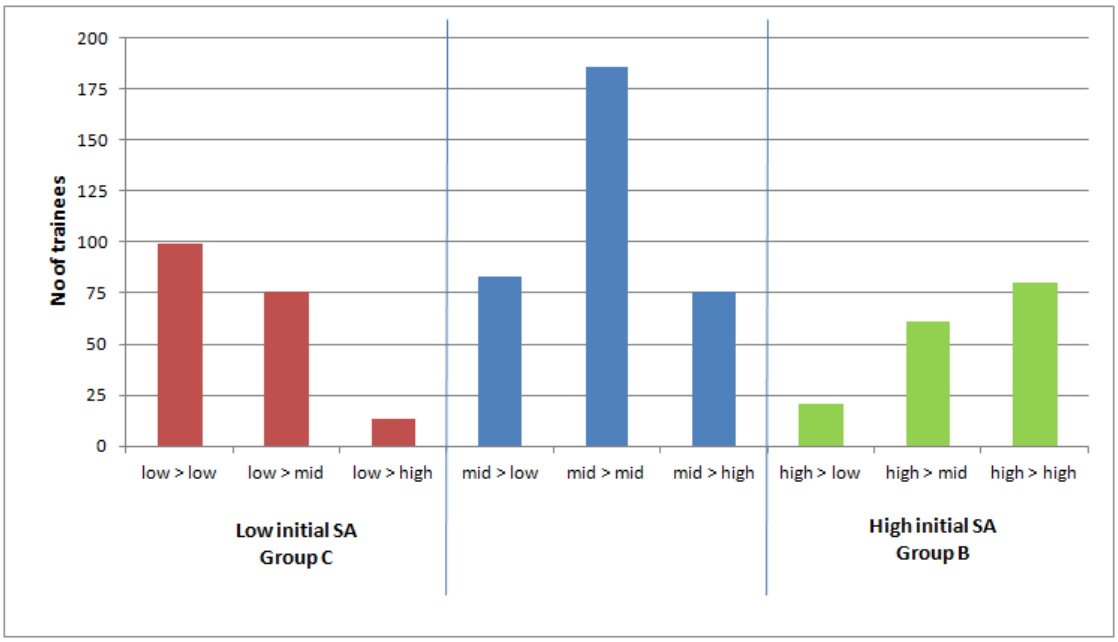

Figure 15. Count of Trainees by Relative Self-Assessment Group in Early and Late Periods. Trainees assigned to the low self assessment quartile (Group C) are in red, while those assigned to the high self assessment quartile (Group B) are in green.

In order to gain a better *a priori* understanding of the development of self-assessment aptitude between the beginning and end of their first training year a group of trainees was identified who had at least one comment of self MSFs in both posts one and three – GROUP H (trainees who commented on their self-assessments in both posts 1 and 3). There was no requirement to comment on MSF forms and a minority of trainees did so at least once in both posts (Figure 16).

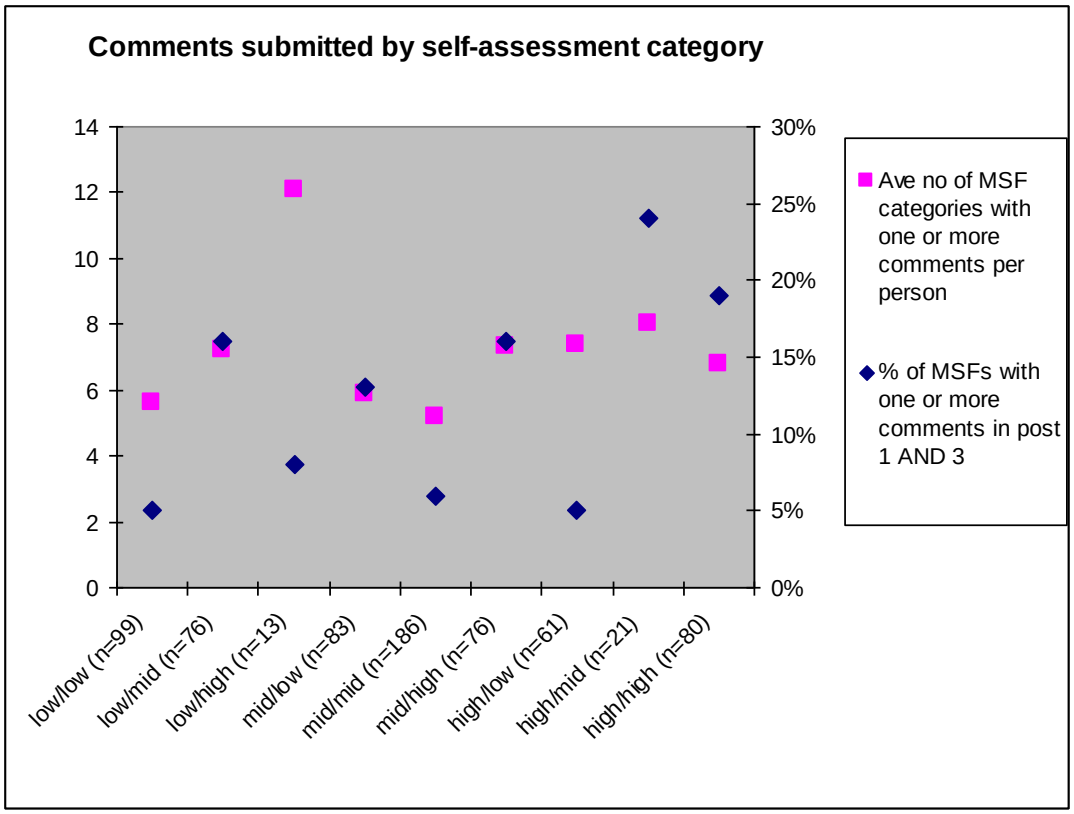

Figure 16. Number of Comments Among Self-Assessors by Category (1st and 3rd post) and the percentage of total submitted MSFs per trainee

Figure 16 shows the average number of MSF categories in which trainees from each self-assessment group included comments (pink square). The blue diamond indicates the percentage of MSFs which comments in the periods under study. As shown in Figure 16, there is no obvious pattern among the number of comments submitted in relation to self-assessment category in this dataset.

5.3.4 Textual Analysis of Subset Comments (Group H)

All comments for trainees who submitted self-assessment comments for posts one and three were subjected to a detailed thematic analysis (Group H, Figure 14). Initially NVivo (a qualitative analysis package) was considered to do the analysis, but it was more practical for the text to be extracted into a spread sheet and reviewed in detail, identifying all coherent issues expressed. Seventeen distinct themes were apparent

and a matrix was constructed to map the nine categories of self-scoring against these themes.

5.3.4.1 Perceptions of Improvement

Nearly all groups, but particularly those that began and ended their training year in the mid groups (mid/mid) commented on their own improvement, with only the high/high group not commenting on any self-improvement. “I feel that my practical skills and clinical judgement have greatly improved over the past year” was a typical comment from a mid/mid trainee.

Many commented that they felt they still needed to improve their skills. “I'm still far from happy with my ability to formulate management plans independently, but I do feel this skill is developing with continuing experience”, noted a low/low trainee. Again these ty comments were clustered within those who started in the low or mid groups, with only one of comments from a high self-assessing trainee who appeared in them self as mid in the third post.

There were a number of comments about the ongoing need to “learn”, rather than the more general “improve”, though these shared characteristics with those above. Trainees citing their need to learn appeared more often in the low or medium groups initially; only one high scoring post trainee cited this, who then self-rated in the low category in post three (suggesting the recalibration effect). Interestingly, the majority of these comments also fell in the clinical skills categories; in line with the literature that self-assessment of these skills is more accurate.

5.3.4.2 Self-Doubt

Comments in which trainees registered doubt in their own abilities were found across the self-assessment categories, but were most concentrated among those who rated themselves the lowest in both posts (low/low). “I still sometimes struggle to explain things to patients particularly if I'm not sure of things myself”, is a typical comment in that the trainee expresses self-doubt but goes on to say “I do, however, ask if patients/parents have any questions. If I can't answer/explain something satisfactorily

then I involve a senior who can".

There were also many comments that went beyond doubt expressing genuine fears about their work. "I do sometimes panic in the acute setting" relates a typical comment, though these stronger comments are always qualified with reassurance that help is available when required.

Numerous trainees commented they wanted more experience to feel more confident. Again, these self-comments came overwhelmingly in the clinical skills categories, with the only comment in the non-clinical skills being with a non-native English speaker wanting more language experience for patient encounters.

Amongst the low/low group in particular, there was doubt and concern about being "not quick enough"/ "too slow", with comments often linked to reported lack of confidence and/or knowledge. Amongst in the self-assessment high group initially there was a single comment relating how quickly a trainee thought they were able to carry out their duties; however, by the third post this trainee scored themselves as low, reported that they now "try to make time for my patients" and wrote of how much they had learned, and still had to learn.

5.3.4.3 Awareness of Self

Although the comments registered on the ePortfolio were usually brief (mostly less than thirty words), there was still sufficient detail to identify where trainees expressed self-ratings with awareness of the skill levels of their peers. These comments appeared in the low/low, low/mid, mid/low and high/low categories. "I believe my knowledge is at a similar level to that of my colleagues, but I still feel this is one of my main areas to work on", comments one low/low trainee, a sentiment that is echoed in all four previously mentioned categories. Similarly representative is the mid/low comment that "I feel that it is essential that I know my own limitations" in that these trainees' comments state or allude to awareness of self in relation to others.

It is notable that only high (first post) self-raters recalibrated themselves downwards by the third post. This suggests they were able to so in with insight of where their abilities fit in relation to their peers that they gained over the training year, and would

be in line with the established literature. These trainees most often commented on the clinical skills areas as well, e.g. “My fellow FY1s have commented on the neatness and conciseness of my clinical note-taking” again suggesting these skills are more objectively self-assessed. They also were very specific in their comments. Whilst most comments spoke generally and sometimes even repeated the subject area (such as management of acute conditions) verbatim, these self-observations detailed skills such as “intravenous fluid administration” rather than just say “I improved my acute management skills”.

5.3.4.4 Relationship to Others

There were a large number of comments that on the theme of seeking help. These were found across the categories, with slightly more in two improving groups (low/mid, mid/high). Again, there is no strong link but it seems a reasonable presumption that assistance and feedback during requests for help, better enabled trainees to assess their own abilities. “I am confident I know when to call for help, and feel that I am gaining experience in this area by observing my senior colleagues” observed one first post trainee commenting on their acute assessment skills, who in the third post demonstrated more confidence: “Through experience I no longer need to ask for help with everything, but can still recognise when I need input from my seniors.”

Another frequently commented theme was appreciation of the wider clinical team, though these comments were more pronounced in trainees whose self-assessments dropped (or remained low/low) between posts one and three. Remarks such as “I greatly appreciate the guidance seniors in the team give”, “not only doctors and nurses but OTs, physios and dieticians as well - They bring a whole new dimension to patient care” and “Every member of the team has a crucial role to play and I aim to work closely alongside all of them to best treat patients” illustrate the positive response trainees had for the other health professionals. There is some evidence for a reasonable link between the less confident, and perhaps more accurately self-assessing trainee.

5.3.4.5 Expressions of Confidence

Every self-assessment trainee category included individuals that reported unqualified positive assessments of their own ability. Very common were comments that reiterated the subject area they were reporting on, for example for “The doctor is polite to patients” a typical response was “I always do my best to be polite and courteous to patients - there is never a situation where one shouldn't be, no matter how angry or upset they make you.” It is not difficult to imagine new trainees using an unfamiliar electronic system wanting to portray their reported ability in a good light, but it is striking that although all self-assessment categories had trainees who commented in this way, the number of trainees and comments of this nature disproportionally fell in the high/medium and high/high categories.

One theme, the use of absolute descriptors when trainees assess their own skills, was striking in the way it was distributed. The use, in particular, of terms such as “always” when describing a positive behaviour fell overwhelmingly in the high/high and high/medium categories. This supports the notion that those who rated themselves the highest may be doing so in an unqualified manner. Comments from high/high and low/low (or others that fell between posts) are nearly universally distinguishable, with the latter routinely avoiding absolute descriptors and qualifying positive evaluations of one’s own behaviour.

The high/high (as well as mid/mid) categories also solely exhibited one theme of noting improvement within post one. While all categories had trainees reporting on improvements between the posts, those already noting they were getting better in their first medical rotation fell only within the categories above.

5.4 Educational Logs

The Foundation portfolio contained a formative and non-compulsory section entitled “Educational Log” which recorded learning activity. Trainees were able to enter events, tag them with descriptors (e.g. lecture, tutorial, procedure, etc.) and reflect upon them by entering free text in predefined subject boxes (e.g. Immediate Thought, Future Considerations, etc.). Entries to the Educational Log were automatically dated and

trainees had the option of keeping them private (the default) or sharing them with their supervisor, who could then comment upon them.

The Educational Log was not a mandatory part of the portfolio; however, one type of event, Significant Event Analysis, appeared as an event option within the Educational Log and was a required summative assessment in a separate part of the portfolio.

The BEME self-assessment systematic review did not identify any papers to test of good quality that reported a change in learners' activity as a result of self-assessment intervention, and therefore the question was unanswered. The Educational Log of the Foundation portfolio provided the raw data to examine learners' self-reported activity, and any relation it might have with instances of self-assessment.

The following sections describe these potential relationships.

5.4.1 Number of Entries

The average number of events entered by trainees into the Educational Log during their first year of the Foundation Programme in 2007-08 was 18.1. Trainees with the lowest self-assessment scores in the first post, Group C, entered fewer events (16.0), whilst trainees with the highest initial self-assessments, Group B, entered more than average (20.7). It is not known why this self-assessment group engaged more with this optional portion of the ePortfolio, but perhaps they felt it gave them the opportunity to demonstrate a higher than average activity and/or proficiency. The subset of 30 highest self-assessors (Group F) who were rated as low by their supervisors entered slightly fewer events on average (17.2).

5.4.2 Type of Entries

There was considerable variation in the type of log entry that trainees chose to enter.

Across all of the first year of foundation, lectures, tutorials and procedures were roughly equal as the most common type of entry. Readings, courses and exams were by far the least frequent. But variation was apparent amongst other entry types, particularly when the smaller groups were examined, which was perhaps evident as

they were more representative of the behaviours described by the Kruger and Dunning (1999) quartiles.

Between the low and high self-assessing quartiles, the differences were minimal.

Across first year, the most common types of self-reported educational activity were tutorials (24.7%), procedures (23.1%) and lectures (22.5%), and the least common were reading a paper (1.1%) and exams (0.1%). Comparing the lowest self-assessing quartile (Group E) and the highest (Group D) with the entire year does not reveal large differences. High early self-assessors less commonly recorded lectures (18.0%) but more often described an event as "other" (12.2% compared to 8.8% overall). Low early self-assessors more often recorded a lecture (26.4%) but slightly less frequently a tutorial (21.5%, compared to 24.7% overall).

However, both the subgroups of 30 trainees whose self-assessments differed from their supervisor's ratings (F and G) diverged significantly from the overall population's average.

Table 14. *Proportion of Educational Log Type Records Submitted by Each Group*

	Group A	Group B	Group C	Group D	Group E	Group F	Group G
n=	781	162	188	30	29	19	11
lecture	22.5%	18.0%	26.4%	17.6%	13.6%	24.1%	10.0%
paper	1.1%	1.6%	1.0%	2.7%	0.0%	1.2%	0.0%
tutorial	24.7%	24.9%	21.5%	26.7%	20.9%	24.6%	8.8%
reading	3.5%	3.7%	3.8%	0.2%	1.5%	0.3%	1.3%
course	2.6%	3.2%	2.4%	2.5%	2.9%	2.0%	5.0%
exam	0.1%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%
presentation	6.4%	6.9%	7.0%	14.0%	9.2%	14.2%	11.3%
SEA	7.2%	6.4%	8.3%	7.0%	12.1%	6.4%	13.8%
other	8.8%	12.2%	7.3%	9.5%	4.8%	8.7%	8.8%
procedure	23.1%	23.0%	22.4%	19.8%	35.2%	18.6%	41.3%

High self-assessors (Group D and F)

High self-assessors recorded fewer lectures (17.6% compared to 22.5% as a whole), though Group F who were reported as below average by supervisors in both posts were more likely to do so (24.1%). Both groups reported giving presentations (14.0/14.2%) more than the average of 6.4%, perhaps reflecting their self-confidence.

High self-assessors less frequently (19.8/18.6%) entered procedures within their educational logs, with 23.1% entries from the entire population being classed as such.

Low self-assessors (Group E and G)

Low self-assessors varied from the overall average to an even greater degree. They recorded lectures less frequently, with only 13.6% of Group E and 10.0% of its subset (Group G) of this; similarly, they reported fewer tutorials (as all FY1 averaged 24.7%), but 20.9% of Group E and a mere 8.8% of the Group G. Both GROUP E and Group G (12.1/13.8%) recorded more mandatory Significant Event Analysis than the average of 7.2%. This potentially revealed a tendency to be more critical of their practice and learning. Interestingly, they also more commonly recorded a presentation (9.2/11.3%, versus 6.4% of the whole). However, this was still less than the high self-assessor subgroups.

Most striking is the variation between the overall averages (Group A) of entries being noted as procedures (23.1%) compared with 35.2% of the (Group E of 30 and 41.3% of the subset of this (Group G). Although these groups entered far fewer items in their educational logs (as reported above), they were heavily disposed towards recording practical skills. Table 14 depicts the percentage each group registered against each type of educational event. Group H is not included as it constituted any trainee who commented in both posts, and was not differentiated by the self-assessment quartiles that were identified for further analysis.

5.4.3 Entries Made Public

The Foundation ePortfolio was designed to encourage reflection, and for this reason certain sections were private to the trainee, rather than shared with their supervisor. In the electronic environment this was achieved with buttons to the trainee the choice about whether to assign the item was to be as “private” (the default) or “shared”. The ePortfolio can “share” data for any pre-defined group or role, but in this example when they changed the status of the item to “shared” it became visible to their supervisors.

Of the more than fourteen thousand Educational Log entries by first year trainees

(Group A) in 07/08, 68.7% were shared with supervisors. Both Group B (66.3%) and Group C (68.0%) self-assessors revealed their entries to their supervisors slightly less frequently than Group A. Within the smaller sub-groups, variations were more pronounced.

The 30 high self-assessors (Group D) shared only 57.2% and the subgroup F 54.8%.This may reflect dissonance between their self-confidence and the other measures of competence.

Table 15. Proportion of Records Made Public by Each Sub-Group

Group	A	low early			high early		
		C	E	G	B	D	F
% Shared	0.687	0.680	0.663	0.850	0.663	0.572	0.548

5.4.4 Educational Supervisor Comments

Upon changing the default type from “private” to “shared”, the trainee’s supervisor automatically received an internal ePortfolio message notifying them the trainee had made an educational log entry available for review and comment. Only 1.6% of supervisors elected to comment on the log entries. Of these comments, 21.2% were of 1-10 words in length, 35.0% were 11-30, and 43.8% were more than 30 words.

The high self-assessing subgroups (were close to the overall average (1.7% and 1.4% for the Group D, Group F), but there was difference in the low self-assessing Group E with 2.9% of the 30 and 3.8% of Group G receiving feedback from their supervisors. It is possible that these trainees developed more of a relationship with their supervisor, or the supervisor made more effort for this group. Nevertheless, the total number remains very small and it was clear that during this training year very few supervisors engaged with what their trainees were recording in the Educational Log. Unfortunately this version of the ePortfolio did not allow for non-routine data comparisons to be readily made and the manual process of examining each supervisor’s page view log to determine how many entries they actually read, was not feasible.

5.4.5 Self-Comments

Trainees nearly always (>99%) commented in some of the free text boxes for an educational event, rather than just log and classify its type. Comments were typically brief, with 67% under50 words, 18% between 50 and 100 words, 15 % 101-200 words and only 26 events were described with more than 200 words.

All Trainees

Almost all records (> 99% of the time) described “What Happened” and 94% described “Where It Was”. Three of the categories had text entered against them in about half of the time: Contributing Factors (50%), What Was Learnt (49%) and Immediate Thought (44%). Trainees wrote in “Thoughts Now” 32% of records, which was intended to be an area for revisiting items and commenting. The least common item was “Future Considerations”, with only 19% of log items having text entered there.

High & Low SA (Groups B and C)

In the highest and lowest self-assessing quartiles, the high self-assessors were 4-10% more likely to comment in all but one of the categories (“Immediate Thoughts”, where low self-assessors commented 0.3% more often).

Commenting from the smaller sub groups (D–G) did vary (though usually much less) from each other and all of the population, but there was no discernible pattern. Similarly, when word count was examined across text entries and between groups, the differences were seemingly random. Although the smallest subgroups F/G (high 5%, low 4%) both entered text in “Thoughts Now” which strongly implies the revisiting of the entry/experience, the numbers were small.

The relatively higher number of words for papers reflected the tendency to copy noteworthy text from published articles, rather than the trainee’s own words which were more prevalent in other types of record. As none of these records (except one Significant Event Analysis per trainee) were assessed as part of their progression, they are an indication of trainees’ personalised use of ePortfolio as a flexible repository for formative learning evidence.

Only about a fifth of entries had included text entered into “Future Considerations” suggesting that trainees did not engage particularly well with planning for their future learning using this tool. The least used text field “Contributing Factors” was filled approximately 5% of the time.

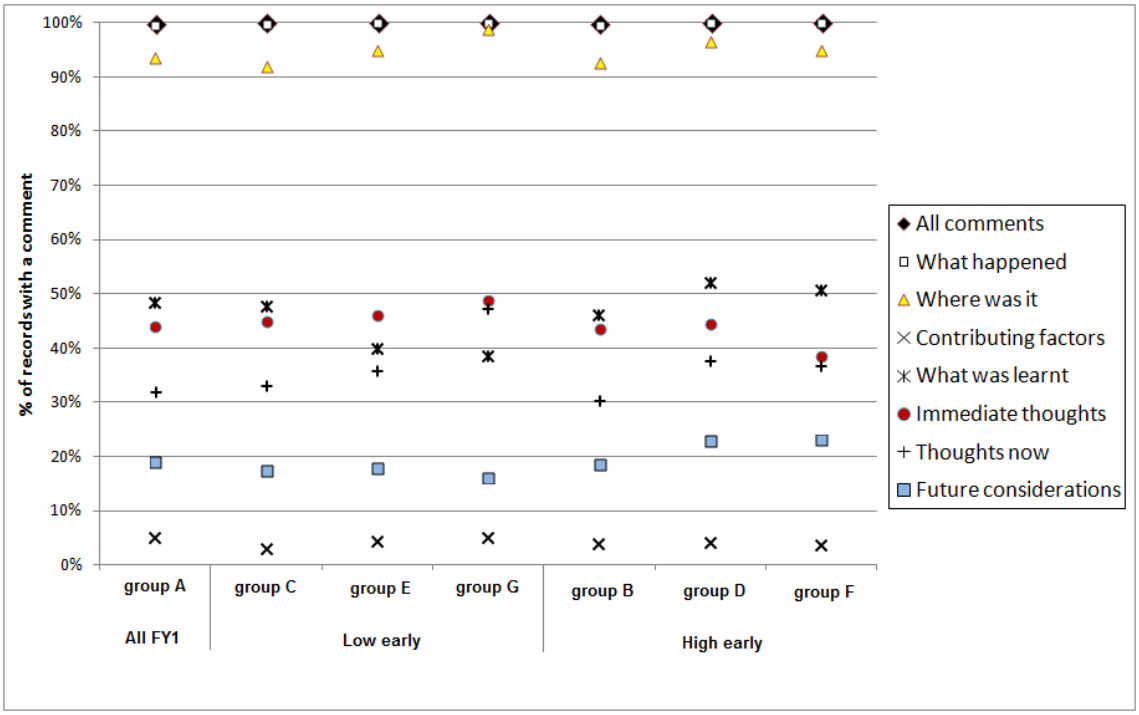

Figure 17. Percentage of Educational Log Records with Comment by Sub-Group and Form Category

5.4.6 Entry Dates

To determine whether assessments, self or supervisor, had a self-reported effect on learning activity as recorded in the educational log dates of assessment and educational log entry were compared. This was done for the two smallest groups (Group F, Group G) who were judged to be most representative of the self-assessment literature by demonstrating potential recalibration.

Low Self-Assessors (Group C)

As discussed previously, the low self-assessors had far fewer (average 7.1) educational

log entries compared with the high group (18.2) and the overall average of 16.0. Of the eleven in this group, four had a single log entry and four more had less than ten. The three with the highest number (10, 16, and 27) were examined in detail. Dates of the self and supervisor assessments were compared with the dates of the educational log entries, and any log entries 5 days or less after assessments were examined in detail.

The trainee with the most entries (27) had only three were recorded within five days of WPA or self-assessments. An examination of the text in the logs gave no indication the assessments prompted planned change in educational activity, but instead they described educational opportunities that arose on the day. A further examination of the remainder of log text found no mention of or even allusion to assessment. The other trainees were similar. The second trainee only had one date match, but the text indicated there was nothing linking the self-assessment and the log entry (the completion of mandatory induction modules). The third trainee had four of their 16 log entries fall within 5 days of completed assessments and again, the date proximities seemed entirely random with the logs describing such things as opportunities presenting themselves to practice procedures such as lumbar puncture. This could be perceived as a lack of engagement and is described in the Discussion.

Results from the three high-self assessing trainees were similar. There were eleven assessment dates predating the forty-three log entries for the first trainee, ten within forty in the second and fifty three within thirty two in the last. However, there was not a single bit of text to imply any log entries were created as a result of assessments; a consequent examination of these trainees' remaining log text (or self-assessments) also failed to reveal active planning of learning activity.

5.5 Improving Perception of Learning Needs (PDP)

5.5.1 Personal Development Plan

The Personal Development Plan (PDP) was designed to set out what the trainee expected to achieve during each placement and throughout the year. It was to be developed between trainee and supervisor, and repeatedly updated when items were

added, altered or achieved. The PDP was the specific tool where learning needs were to be acknowledged and acted upon, and had the potential to inform around the "identification of learning needs" question of the systematic review.

The 2007-08 PDP was however an optional tool and usage was very low. Analysis of the PDP data was further hindered by an incomplete data set – PDP records after 10th February 2008 were unable to be retrieved from the master database. This meant only the data from first post were complete, the second was partial and there was nothing from the final post of Foundation One (see Table 16).

Table 16. PDP Entry Details by Self-Assessment Sub-Group

Group	Post	Individuals with records	% of total	Total entries	Average entries/trainee
Group C	1	61	32.8	217	3.6
Group C	2*	25	13.4	72	2.9
Group B	1	44	27.3	135	3.1
Group B	2*	13	8.1	28	2.2
Group A	1	247	31.6	822	3.3
Group A	2*	78	4.8	186	2.4

** incomplete data set*

Slightly more trainees who initially scored themselves lowest on self-assessments entered PDP items than high self raters (33%/27%), but in both groups fewer than a third of trainees engaged with the tool in the first post. Low early raters (group C) also made slightly more entries (average 3.6/3.1) but again even the groups that did make entries did not do so prolifically. It was more difficult to judge the second post as the final seven weeks of data could not be accessed; however, from the figures up to 10th February, it is reasonable to assume that levels of PDP engagement were set to drop even further.

PDP data for the thirty low self-assessors with high supervisor scores (Group G) and high self-assessors with low supervisor scores (Group F) were examined in detail (also completed for Supervisor's Report in Section 5.6).

Of the 30 low self-assessors, 7 used the PDP entering a total of 18 items. Four of these 7 were in subgroup G where Post Three supervisors also scored them as high; there

were no noticeable differences between the two small groups F and G. Entries were brief and the text revealed little more than lists of specific skills under "Learning Objectives" and infrequent "Action Points" that did not go beyond expressions of commitment to achieve the item(s). No comments referred to assessments or events that provoked a desire to note or change learning needs.

Eleven of the thirty high self-assessors (Group D) entered a total of 29 PDP items; nine of these 11 were in the (sub) Group F where third post supervisors continued to rate them as low. Although these proportions were different than the lower self-rating groups (E and G), it was not possible to draw any conclusions with the paucity of information. Comments made by the Group D were equally scarce and mirrored the contents of the low group.

Items entered into the PDP could be qualified (at point of entry or any time thereafter) as "Completed". Once tagged as complete a message was sent via the ePortfolio messaging system to the supervisor who was then to read the item and further qualify it by indicating it was "Closed". Both the low and high self-assessment groups had similar numbers (44%, 42%) of items in post one being tagged as "Completed" and less (21%, 24%) of items "Closed" by supervisors. Again, this would suggest even those who were motivated to use the PDP did not necessarily follow through on submissions. It should be noted though that completed/closed demarcations would only appear up until 10th February (due to the spilt of the data set) and these may have been noted later in the training year.

5.6 Supervisor's Report

Every Foundation post required the submission of a Supervisor's Report, a structured formal report to record that the appropriate level of competence was achieved during that post. For this study, these reports were analysed for posts one and three of the first year of the Foundation Programme for trainees that self-assessed as either low (Group C) or high (Group B) in their earliest self-assessments. In Group C there were 186 self-assessors in post one with the required Supervisor's Report and a slightly (181) smaller number in post three (from a total of 188). Of the group B high self-

assessors 161 had Supervisor's Reports in the first post and 160 in the third (from a total of 162).

Individuals who initially rated themselves as low but whose supervisors rated them as high in both posts were identified for further comparison, as were high self-raters whose supervisors rated them as low.

Table 17. Supervisor's Report Score Comparison by Sub-Group

Score given by Supervisor	Post 1 low Self-Assessment(Group C)	Post 3 low Self-Assessment	Post 1 high Self-Assessment(Group B)	Post 3 high Self-Assessment
<5	2 (1%)	2 (3%)	1 (1%)	-
5	25 (13%)	16 (9%)	15 (9%)	11 (7%)
6	82 (44%)	84 (46%)	58 (36%)	46 (29%)
7	77 (41%)	79 (44%)	87 (54%)	103 (64%)
Total	**186**	**181**	**161**	**160**

Table 17 shows that Supervisors appeared reluctant to rate trainees lower than five on the seven point scale, which functioned as an unofficial threshold.

According to supervisors global scores high self-assessors demonstrated the greatest improvement between posts. All scores six and under decreased by the third post, with those scored a perfect seven by an early and late supervisor increasing by 10 %.

A substantial difference is evident when low initial self-assessors are compared with high. Low self-assessors are more likely to score six or below, whilst 13% more high self-assessors achieved a global score of seven from their supervisors in the first post. This margin increased even further between low and high self-assessors in last post three, to 20%. Supervisor score seemed to match with trainee's self-assessment which contradicts the literature as described in Section 2.8.7. Supervisor comments are discussed in Section 5.6.3, and the implications are described in the Discussion.

To further compare the groups of self-assessors, an average was taken of all scores for specific competencies, excluding the global, which was a single universal score per MSF. The mean of all scores was slightly lower than the global in three of the four groups (global scores for high self-raters in post one being identical to All Scores means). High self-assessors remained better rated by their supervisors in the non-

global averages than the low self-assessors.

Table 18. Mean scores, All and Mean, by Self-Assessment and Post

	Post 1 low SA	Post 3 low SA	Post 1 high SA	Post 3 high SA
All scores (mean)	6.1	6.2	6.4	6.4
Global (mean)	6.3	6.3	6.4	6.6

To measure how the Foundation trainees' self-assessment scores compared to what is known in the wider research it was then necessary to isolate the low self-assessors who were rated highly by supervisors and, conversely, the high self-assessors who scored the lowest in Supervisors' reports.

For each subgroup, the individuals' scores were compared to how their supervisors rated them in the third post. Whilst the supervisors would always change between posts one and three, this (the supervisor's report) was the only available point of comparison with a consistent measurement that could be applied between the posts.

5.6.1 Low Initial Self-Assessors

From the 186 initial low self-assessors from C, the thirty with the highest average Supervisor's scores (Group G) were selected; the average score for the larger group was 6.1 whilst the average for those in the top thirty was 6.8 or higher.

By post three, eleven of the thirty continued to rate themselves as "low" whilst 17 self-rated as "mid" and only two adjusted their self-score to "high".

Thirty were rated as high by one supervisor, and posts one and three were compared by taking the third post supervisors' scores and examining them for evidence of recalibration. The lowest third (<=6.0) had 9 trainees, the mid (6.1-6.5) 10 and the highest (>=6.6) 11.

Of these 11 trainees who scored as high in both supervisors' reports 6 rated themselves as low in both posts, 3 as low then mid and 2 as low in the first but high by the third.

Table 19. Comparison of Self-Assessment and Supervisor Ratings for High Group

Self assessment (post 1/3)	High SR Post 1	High SR Posts 1 and 3
Low/low	11	6 (55%)
Low/mid	17	3 (27%)
Low/high	2	2 (18%)
Total	**30**	**11**

As those with high supervisor scores in both posts can be assumed to more likely be better trainees, it is notable that approximately half did not recalibrate themselves away from low and just two of the eleven scored themselves in the high category by the end of the training year. All comments from the eleven were analysed for further insight (below).

5.6.2 High Initial Self-Assessors

Of the 161 trainees with high initial self-assessments (Group B), the thirty with the lowest Supervisor Report scores for post one were selected for further investigation (Group F). The average score for the larger group was 6.4 while the average for the subgroup was 5.9 or less.

By the third post, 14 of the thirty high self-raters continued to rate themselves as high, 9 adjusted their self-score to mid-range, whilst 7 went on to score themselves relatively low.

As with the low self-assessors group, supervisors' scores in post three were checked to determine whether the post three supervisor's report also scored them relatively low. The scores of the larger group of 161 were separated into thirds with high scores being 6.8 or more, mid scores 6.4-6.7 and low scores being equal to or less than 6.3 – clearly these are closely demarcated groups and therefore only tentative conclusions should be drawn.

Of these 19 trainees who scored low in both supervisors' reports, 5 self-rated as high/low, 7 adjusted their scores from high to mid and 7 self-assessed as high in both posts.

During the audit, ePortfolio records for around 180 first year posts were identified as

having a missing mandatory form (Supervisor's Report or Certificate of Performance). The reasons for the missing reports were investigated by direct contact with local administrators. The most common reason (in about half of cases) given was a local policy that deemed the post not to require a formal sign off. Other reasons given were that the trainee had resigned, trainee absence, IT difficulties and technical issues with the status of the post. In around 50 cases the Deanery was not able to explain the reason for the lack of the mandatory form, most of whom involved successfully completion. Just more than one quarter of these trainees adjusted their self-ratings to match their supervisors', while less of the former (low early self-assessors) group (18%) did so. Similarly, a higher proportion of self-raters did not recalibrate to match supervisor's ratings from the low early assessors group when compared to the high (55% to 37%). This conflicts with accepted findings in the literature in that one would expect low early raters of ability that were highly rated by externals to be better able to see their abilities in perspective. Those with less ability but more (misplaced) self-confidence among (the high early Group F) were postulated to be less likely to readjust their self-ratings unless their actual skills or knowledge improve first. However, as the overall average supervisor report scores (global scores excluded) remained at 6.4 for both posts there is no evidence of this in these trainees.

5.6.3 Textual Analysis

Within the 2007-08 training year it was not uncommon for little or no text to be entered with the submission of mandatory Supervisor Reports. Between the two groups of low early and high early self-assessors who had consistent (high or low) supervisor scores between first and third post supervisor reports (F and G) there were thirty individuals (11 low early and 19 high early). These thirty trainees worked in sixty posts of which twenty-six the Supervisors Reports contained no comments. Six of the thirty had no comments whatsoever registered for either post in the supervisors' reports for this training year. Nevertheless, despite the paucity of recorded text, the available comments do give insight into both groups.

GROUP G (low self-assessment, opposing supervisor rating)

Amongst the supervisor comments for the low early self-assessors (high supervisor ratings, Group G), there was not a single negative, or even neutral comment. When comments were brief, they most frequently referred to specific skills or made generalisations that trainees were “excellent”, “very good” or “highly competent”. Other affirmative supervisor comments were attributed to communication, compassion, accuracy, team-playing etc., and again were without qualification, despite the fact that the trainees initially (or even in both posts) rated themselves low compared to their peers.

Four of the eleven low early group had comments where supervisors judged the trainees in relation to their peers and/or developmental phase. “Excellent - well beyond expectation for this stage.”, “a first class doctor who others (F1 trainees) respect and rely upon” and “Outstanding, one of the best trainees we have had” were all comments supervisors gave trainees who rated themselves as low in both first and third posts. The first two comments were registered in post one, the third in post three.

The final trainee that had comments attributed to their ability relative to their experience/peers was a first post comment in which the supervisor related the excellence of aptitude, but went on to say the trainee should have “more self-belief and confidence in [their] own abilities”. There was no comment in the third post, but in the latter post the trainee self-scored in the high category, some evidence of the recalibration effect.

GROUP F (high self-assessment, opposing supervisor rating)

The comments attributed to the high early self-scoring trainees (scored low by supervisors) were more complex, frequently attaching caveats to noted improvement or qualifying negative comments with potential positive change. Unlike the above group, these trainees frequently had comments noting they had learned a lot during respective posts.

The trainee with the lowest post one average score (4.6, group average 6.4) was

commented in post one as “need(ing) to be more involved. This should come with familiarity with UK medicine”; by post three they averaged 6.0 (group remains at 6.4) and the supervisor noted the “communication improvement” with the “steady (overall) improvement.” This trainee self-rated low in the third post, perhaps becoming aware of their shortcomings compared to peers, regardless of the ongoing improvement.

Another trainee that adjusted their self-rating from high to low between the posts, was described as “initially immature, but got better” and as having “organisation skills that impaired patient care, especially at the beginning”. No comments were entered for their third post, but these follow the pattern of supervisors balancing criticism with improvement, and the trainee themselves recognising shortcomings over time.

It is entirely possible that supervisors were reluctant to give negative comments. One trainee (self-rating high/mid) was described as “adapting well” and “progressing” with an average of 5.4 (group 6.4) in post one. The third post supervisor notes they are “probably about average” despite still scoring 5.4 in the last post.

Another self-rating high/mid trainee had numerous problems with communication and organisation but each was tempered with notes of improvement. Unfortunately the third post supervisor neglected to comment, but the average of scores actually fell between the posts (5.6 – 5.0).

There were occasions where ability was seen to decline between posts by both trainee (high to medium) and supervisors, notably commented for one in a case of personal bereavement.

Comments on those who self-rated as high in both posts were similar in that they qualified criticism and nearly always spoke of improvement. A trainee whose average scores (by supervisor) ranked at the bottom in both posts (5.0/5.0) has post one comments that spoke of “initial problems” yet went on to say they were “satisfactory for this stage”. It is difficult to see how the worst rated trainees are still gauged to be satisfactory, but possibly alludes to a culture that is reticent to note poorer performance. By post three, they are seen as “functioning effectively” and “improving”, despite no improvements in the average of their scores. This reinforces

the concept adopted in this thesis that the scores submitted were a relatively blunt measure, whereas comments added richness and depth to the understanding of the individual's training.

Another high/high self-rater needed "experience", "improvement" and "more progress", and repeatedly told to "get help if you feel under-confident". It is questionable if confidence was the issue, as the high self-ratings would attest to – a finding consistent with the wider literature that poorly performing individually need to improve their skills before they can be aware of their deficits. Although this trainee's average supervisor scores improved from 5.2 to 5.8 between posts one and three, her post three supervisor worryingly comments they had "not seen her" perform many of the skills, "but others say she's competent".

Other high/high self raters had similarly negative comments or low-end scores by supervisors, but these were qualified or not seen as impediments. A supervisor lists five separate areas (including organisation, communication and team-working) needing improvement for one trainee to become "more effective" but then comments that these are "not serious issues" and the trainee "should succeed". In post three the trainee still has below average scores (5.8/5.9) and the third supervisor notes "well intentioned but abrupt", "rather esoteric in differential diagnosis" and "little experience" but qualifies with support for their attaining MRCP and their "gaining self-reliance".

A trainee scoring below average (5.8) in post three is described as a "good doctor" with an "appropriate level of experience" and "no issues to impede progress". The suggestion is that unless there are exceptional circumstances, all trainees are expected to progress without reservation.

5.7 Record of Progression

At the end of each post supervisors were required to submit two records to indicate the trainee's competence. A total of 224 (first year) and 241 (second year) posts did not have a submitted Supervisor's Report; 9% and 11% respectively (12% and 16% for Certificates of Performance – which was a simple single form to indicate competence

and has since been discontinued as an unnecessary additional step). For a small proportion of these missing reports (7%) the reason cited was a performance issue known to the Deanery relating to seven first year and eleven second year trainees (indicating that underperforming trainees tended to have more than one missing report). Other reasons for failure to submit a mandatory report are shown in Table 20.

Table 20. Distribution of Reasons for Non-Submission of Supervisor's Report or Certificate of Performance

Reason for non-submission of Supervisor's Report or Certificate of Performance	No of first /second year trainees (% of all trainees)
trainee resigned	1 / 5 (1%)
legitimate absence	6 / 13 (1%)
post reported not to require sign-off*	74 / 0 (5%)
persistent difficulties with IT equipment	11 / 6 (1%)
technical issue with status of trainee or post †	23 / 17 (2%)
no specific reason known to Deanery - trainee passed ‡	41 / 107 (9%)
Deanery unable to provide reliable information	27 / 53 (5%)

** mostly in two-month posts*
† including test trainee IDs entered for training purposes
‡ including a region's decision not to submit both SR and COP, or to accept a paper copy

Very few reports were submitted with a low overall assessment score for Supervisor's Reports (n=17 first year, 16 second year) or failure to achieve competence for Certificates of Performance (n=13, 19 respectively). By checking the content of associated comments, and discussion with Deanery administrators it was apparent that in several cases these were again erroneous scores entered by supervisors and therefore not indicative of a competence or professional issue. Using the completion data which was made available, it was estimated that less than half of "unsatisfactory" reports coincided with a trainee who did not achieve competence.

5.8 Programme Completion Rates

Details of trainee programme completion were not recorded electronically in the ePortfolio due to the required (paper-based) format for reporting to the General Medical Council. Separately collected data indicated 99% of trainees completed the

training year satisfactorily. This correlates well with those identified through the ePortfolio as having missed mandatory elements (supervisors report and certificates of competence) without a known local reason (0.9% and 1.4% of first and second year trainees).

5.9 SUMMARY

The purpose of this case study was to evaluate the self-assessment systematic review's core questions, and test the hypothesis that the Foundation group would adhere to the behaviour and findings in the wider published evidence. The ePortfolio was effective in extracting the required data, albeit with some considerable effort, and it provided mixed results. Some of the findings were confirmed, but often there was a paucity of information making comparison impossible.

There have been substantial changes to the Foundation ePortfolio since 2007-08, and the understanding and acceptance of Foundation's changes is now much more ingrained. A consequent analysis of these same questions would likely result in a much more detailed picture.

Summary Points

- Analysis of the data from this training year quickly revealed that in practice the assessment and educational processes varied greatly from what was intended, notably in terms of engagement and following chronological processes.
- Quartiles as defined by Kruger and Dunning and replicated widely since were discernable within the population after weighted ranking of scores.
- Qualitative analysis of text entries of self-assessors who commented in first and final training posts broadly reflected what was found in the wider literature in categories such as expressions of confidence.
- Both low and high self-assessors demonstrated behaviour consistent with the wider literature with regards to their learning needs perception (Educational Log), but this was not universal.
- Analysis of use of the PDP to determine the perception of learning needs was inconclusive due to lack of widespread user engagement with this non-mandatory component.

  Analysis of the Supervisor's Report to evaluate improvements in clinical practice within self-assessment groups found contradictions with the literature in terms of supervisor-assigned scores, but confirmation of published findings when their comments were analysed.

6 DISCUSSION

6.1 INTRODUCTION

The following chapter discusses the testing of the self-assessment review's questions in the case study of Foundation training, how the findings of the portfolio review compare with the use of the ePortfolio in Chapter 5, and the issues in a wider educational context.

The professional, as an individual, has historically held the responsibility for evaluating and maintaining their own competence. Whilst the merits of this approach are debatable, the task of monitoring and maintaining one's own professional standing is compounded yet facilitated by rapid technological change and unprecedented growth of information. At the same time the process of self-monitoring has become far more formalised and transparent across the health professions in recent years (Mann 2011, Epstein et al.,, 2008). This research evaluated the inherent challenges of self-assessment within the health professions and how the constantly evolving technological advances could potentially aid the process, in part through an electronic portfolio.

Despite its increasing prevalence, evidence to support the effectiveness of self-assessment is at best sparse. Numerous surveys and studies confirm that the majority of people perceive themselves as being better than average across wide ranging activities. Kruger and Dunning (1999) synthesised this evidence in experiments ranging from logical problems to judgements of one's own humour. Their results concluded that not only were the poorest performers unaware of their incompetence but they could not distinguish good performance in others for precisely the same reason – they lacked the knowledge about the skill itself. This "perceptual deficit" means poor performers simply do not know what good performance looks like. The expertise required to perform a task well is the same expertise that is required to recognise good performance in others. For this reason, the top performers in an area (who tend to initially under self-rate) are able to recalibrate more accurately when exposed to their peers' performance.

The sections below outline and discuss the representativeness of the case study's population. Beginning with an examination of the population of the case study, and importantly whether it breaks down into the seminal quartiles reported by Kruger and Dunning (1999), this section details how the findings of the two reviews – both confirmations of and gaps within the evidence – relate to the population. Firstly the component parts affiliated with the self-assessment systematic review's questions are discussed, followed by issues of the portfolio review, and then more detailed discussion on relevant related issues of self-assessment in postgraduate healthcare training. The discussion concludes with an examination of the potential directions of future research and development for both self-assessment and its use in electronic portfolios.

6.2 Population

Access to the NES ePortfolio enabled an analysis of trainee assessment data that was without parallel. Although Foundation had been recently implemented and was undergoing continual change, the comprehensive scope and detail of the data that was electronically collected provided an opportunity to test the effectiveness of self-assessment in healthcare education, whilst simultaneously being informed by the findings of the portfolio review. It was also an opportunity to examine a recommendation of the systematic review, to explore the possible cognitive pathways between self-assessment and professional performance.

This was a retrospective analysis of data, and the potential bias and confounding factors inherent in this type of analysis have to be acknowledged. There was also a lack of demographic data on the users, as the ePortfolio was not intended to (and still does not) capture information about gender, age, race, etc. This was an opportunistic analysis of the data available and the educational tools selected for the programme. In fact, these tools have to be acknowledged as having their own limitations in terms of validation, which will be noted in the discussion below.

The selection of such a group is likely to involve some unavoidable sampling bias as, by definition, it excludes trainees who were less inclined to add their reflections on their

own performance consistently at the start and end of the year, and who therefore may have spent less time considering their self-assessment. However it was done with the aim of fully exploring the practice of self-assessment using all available data.

6.2.1 Quartiles

The initial analysis of the Foundation data was to determine if the population quartiles widely described elsewhere in the literature could be reliably identified in the ePortfolio trainees – and crucially whether the lack of insight of the lowest performers would be replicated in this data. Multi-source feedback (MSF) required of Foundation trainees included mandatory self-assessment and this was used to identify high and low quartiles first described by Kruger and Dunning (1999) and described in Section 2.2. The data was then analysed against three of the four central questions of the self-assessment review, as each question had a correlating section within the ePortfolio to provide a measure. Examining the impact on patient outcomes was not possible as the ePortfolio does not contain patient identifiable information as per the Caldicott Principles ("Personally identifiable information items should not be used unless there is no alternative."), and indirect measures were not practical or possible.

Importantly, because there was in fact a narrow range of differentiation in both self and non-self-assessment scores it was problematic but not impossible to define quartiles by weighting scores (Section 5.2). The groups did then fall into manageable subgroups for analysis, but it is noteworthy that in the Foundation data these were not readily identifiable as stated or suggested in other studies with positive skewed distributions (Edwards et al., 2003, Ehrlinger et al., 2003, Lane and Gottlieb 2004, Langendyk 2006, Ehrlinger et al., 2008). Anecdotally, it would appear that there was an unofficial practice of scoring trainees within this narrow range and/or not registering assessments until it was felt that the trainee merited a higher score. The methodology employed was developed iteratively by taking into consideration the initial findings of skewed data. With a different dataset, more quantitative and statistical analysis may have been possible, allowing cross-validation of multiple assessment tools or other quantitative data. Movement between the first and final post of Foundation Year One

was in proportion to other studies and gave some confidence that these trainee assessments scores would be similar to previous studies (with some evidence of trainees recalibrating between opposite ends of the scoring).

Although the ability to comment against assessment scores (self and non-self) was possible, a minority of raters chose to do so. Nevertheless, the thematic examination of the comments found strong correlation with themes in the wider literature following on from Kruger and Dunning (1999)(Edwards et al., 2003, Ehrlinger et al., 2003, Hodges et al., 2001, Lane et al.,2004, Mandel et al., 2005).

Perceptions of self-improvement were noted in only the low or mid groups, suggesting they were more aware of their ability and saw the need to comment on it; these comments were largely in the clinical skills domain, which again matches expectations that tangible skills are more accurately self-assessed.

Similarly, self-doubt and lack of confidence was mostly expressed by low self-raters commenting they need to improve in clinical skills areas. One initially high self-rating (but low by peers) trainee explicitly described recalibrating themselves as lower after observing others' clinical skills. The established literature demonstrates this is far more likely to happen in reverse, and although only a single anecdote, it is a note that successful recalibration of self-assessments of clinical skills does happen by low self-raters.

It was rare however for trainees to write of their skills in relation to other trainees, primarily skills comparison was noted by low self-raters and it was always for clinical rather than "soft" skills. It was striking that there was no mention by any trainees of how well they thought they communicated with patients, strongly suggesting that this area of self-examination was neglected and not encouraged in their training. This corresponds with other studies (Millis et al.,2002) that described trainees having difficulty judging how well they communicate with patients, and the fact they rarely receive feedback on their patient communication skills. This does not bode well for self-assessment accuracy as the evidence strongly suggests self-assessment in this area can only really achieve its potential accuracy in conjunction with other assessments and awareness of the abilities of peers.

Certain trainees (low/low self-raters, and those whose self-rating dropped between posts) did frequently mention their peers and the wider clinical team with appreciation. This demonstrated an awareness of the importance of others' roles in successfully learning in the workplace – an awareness not mentioned by any high self-raters.

High self-raters distinguished themselves in comments from mid and low raters by far more frequently expressing their confidence, use of absolute descriptors (e.g. "I *always* make certain I do this to the highest standard") and were the only group(s) to note improvements in their own abilities within their first post practicing medicine.

Broadly, the text entered within self-assessment MSFs is very much what would be expected of each quartile and gave confidence that the groupings would go on to mirror expected behaviours across other ePortfolio activities. However, it must also be noted that commented self-assessments were in the minority and frequently lacked a sufficient level of detail for analysis. Potential reasons for this are discussed in detail in the sections below.

Whilst the population would be broadly broken down into the theorised quartiles, the behaviour the subgroups demonstrated only weakly corresponded to existing findings after qualitative analysis sometimes elicited conflicting or ambiguous evidence. The sections below explore the issues involved with the medium (ePortfolio) and how self-assessment was implemented and conducted as a core component of Foundation doctors' education.

6.3 Self-Assessment

The work of the BEME Collaboration has confirmed that doing a systematic review of the education literature is indeed different from doing one focused on a clinical question (Hammick & Haig, 2007). The disparate types of evidence that need to be considered in educational research, as well as all the confounding factors that can influence learners, mean that a formal synthesis (let alone a meta-analysis) is not possible. The poor methodological quality of studies (common issues included unsustainable assumptions, data omissions, and questionable generalisability) also

contributed to the difficulties of conducting a formal systematic review. However the rigorous nature of the trawl and review of evidence supports firm conclusions on an albeit small evidence base.

Nevertheless, the BEME reviews were able to comprehensively retrieve all relevant evidence, appraise it objectively against agreed frameworks, and arrive at a consensus in a transparent and reproducible manner. The systematic review into the effectiveness of self-assessment was, however, unable to answer its specified original questions, which were in turn examined in the case study:

- Does (self-assessment) improve the accuracy of learner perception of their learning needs?
- Promote an appropriate change in learner learning activity?
- Improve clinical practice?

The review employed rigorous methods and analysed as comprehensive a search of the evidence base as was possible, but both the quality of the published (and un-published) papers was frequently less than sufficient for inclusion. The review concluded self-assessment was difficult to define and more difficult therefore to objectively measure.

Nevertheless, the review did find some positive evidence to answer subsidiary subjects which were in turn examined in the case study), specifically: there was multiple health sector-specific confirmation of Kruger and Dunning's seminal work (1999); peer-assessment was shown to be more accurate than self-assessment (and could be used to validate the latter); practical skills appeared to be more readily and accurately self-assessed than soft skills; and benchmarking and feedback appeared to improve self-assessment accuracy. The Foundation data presented an opportunity to test this.

Other subsidiary results were inconclusive: although gender differences were frequently examined the evidence was equivocal, culture and race were seen to have no impact on self-assessment ability, and that the suitability and acceptance of self-assessment to both learners and teachers is very seldom considered. These areas could be very readily explored if this data was included on ePortfolio, but existing

governance continues to preclude the inclusion of these data fields.

It should be noted that the ePortfolio does not identify the gender, race or culture of trainees and the inconclusive results of the self-assessment review, as well as the weak evidence in the portfolio review that females engage more with a portfolio, could not be tested.

This review identified the pressing need to examine self-assessment data within the place of learning environment, rather than the context of external and/or disassociated skills. The use of self-assessment is still widely prevalent and growing, in both summative and formative assessment, and it is increasingly forming part of the decision making process in high stakes environments such as registration with regulatory bodies and re-certification.

Despite the widespread and growing use of self-assessment it is notable there were no studies that focused on the opinions of those undertaking the self-assessment towards the activity. Some studies acknowledge that it can be threatening and stressful, but rather than exploring the attitudes and perceptions to self-assessment there is a general assumption that users find it acceptable. The literature describes the importance of a well-considered and run implementation process to maximise the potential of self-assessment (Crawford 1998), as well as the use of portfolios (Tosh 2005), but there was very little mention of either happening in practice.

Similarly, in Scottish Foundation training, there was no formal introduction of self-assessment within the training year, just the assumption that the ends justified the means. Engaging the users in the process and its principles may well have significant impact on self-assessment, but a combination of lack of time and resource, as well as the what is often described as a “top-down” approach to implementing educational interventions, has meant it could be argued that this has not occurred in any meaningful way.

There are a variety of factors that will ensure that the avocation of self-assessment continues. It is increasingly seen as a cornerstone to professions, and one that any competent individual professional should and could do. Practical considerations are featured heavily. Peer and other types of assessment are more time and resource

intensive than when evaluating one's self and in a time of diminished resource it can only be expected that organisations and professional bodies will look to self-assessment's potential to ensure the quality of practice.

Self-assessment is not a succinct or transferable skill, but one that is innately connected to the particular skill or situation being assessed. Good or poor self-assessment cannot be generalised. Lack of insight cannot easily be tested for and therefore is most needed where it is least available.

Portfolios, particularly e-portfolios, are an ideal medium to test such divisive opinions such as the potential to enhance self-assessment through triangulation with other assessments, in that they can instantly display and compare all the data they collate. There is a general consensus, particularly in the medical and dental literature, that self-assessment would produce more stable, reproducible and accurate measures when used in conjunction with other methods (Rees 2005). E-portfolios themselves offer particular opportunities as they ideally operate in tandem with e-learning environments, or supporting e-learning content themselves.

The increasing pervasiveness of self-assessment continues however, and the changing technology of the learning environment is providing opportunities that have not been previously possible.

6.3.1 Educational Log: Does (self-assessment) Improve the Accuracy of Learner Perception of their Learning Needs?

The ePortfolio's Educational Log was the learner's record of activity. Similar to the PDP, this was deemed a "mandatory" component in the 2007-08 training year, but with ambiguously defined requirements of use (except the peer review of one SEA) and no stated penalty for not engaging with it (there was an assumption the Educational Supervisor would ensure it was used). Although engagement with the Educational Log was greater than with the PDP, it was still sporadic and moderate. The variation is noteworthy, and contradicts other study findings that the high self-assessing quartile would be less educationally engaged and the low quartile more.

Nearly three quarters of all Educational Log events entered by Foundation Year One

(Group A) trainees were one of three types: lectures, tutorials or procedures. As each narrower/smaller sub group is examined there is increasing variation in Log entries from Group A, as well as between the high and low self raters. Both strands (high and low self-assessors) were designed to test sub populations of Foundation trainees against expectations from the published literature. Further exploration to tease out extreme behaviour was done by selecting the highest and lowest scoring 10% of trainees and identifying among them, the group with strongest variation from their supervisor's opinion of them. Among these extreme sub-groups some distinctions were notable. Low self-assessors who were regarded as above average by their supervisors (Group G) were less likely to record activity in their Educational Log and when they did it tended to be practical Procedures. Conversely, the opposite group (F) entered fewer practical procedures and recorded their own Presentations as events more than any other group. This could be viewed as demonstrating that Group G focused on what was demonstrable and practical, where the more confident Group F was more likely to record (and possibly do) presentations.

Group G also stood out noticeably from their peers in other ways. These trainees entered a smaller average number of events which contradicts any assumption that the most competent trainees are more likely to engage with the recording and sharing of educational events. It also demonstrates that the lowest self-assessors are far less likely to engage with their educational logs, possibly because they were less likely to spend time with an optional recording of learning and sought to remedy their perceived short-comings with experience.

Trainees varied with regards to how many entries they made public as well. Overall, 69% of entries were shared with supervisors, with Group E slightly less frequently (66%) sharing their entries, perhaps indicating a lack of self-confidence of low self-assessors; however, the Group G differs by sharing 85% of entries. This smaller group of doctors, who were rated highly by both supervisors, demonstrate both candid and more accurate assessment of self and openness with regards to their thoughts and activities, corresponding well with other studies. High self-raters (B, D, F) were increasingly less likely to share entries, possibly illustrating they did not value the

potential for dialogue or criticism as much, but certainly revealing a dissonance between self-confidence and other measures.

Further analysis of the Educational Logs unfortunately revealed very little. Educational Supervisors were very unlikely to comment on Events shared with them, suggesting a lack of time or engagement on their part. Low self-raters (Groups E and G, 3 & 4%) received more feedback on Events than the population average. Possibly this shows more engagements between supervisor and trainee, but given the number of actual comments this is hardly a certainty. High self-raters' scores did not exhibit any trends.

The entry dates of log items were examined to see if there was any relation between MSF scores (self or other) and events recorded. The results demonstrated there was no observable pattern of MSFs triggering activity in the log, and no significant variation between the groups between MSF and Educational Log date entries. Foundation intended the educational assessments and events to exist holistically in the ePortfolio, but this certainly was not what happened in practice in this case. Some of this dissonance could be put down to time elapsed between the activity and its recording in ePortfolio, but there is no way of knowing the extent of this and the elapsed time, it could easily be argued, would have an impact on the accuracy and value of the actions.

Finally, the subject areas where trainees did enter comment were examined across the groups and found large variations in practice. There was however consistent variation between groups in the commented areas with the factual (*What/Where (it) Happened section)* nearly always described, *Immediate Thoughts* and *What Was Learnt* sections appearing slightly less than half the time and *Thoughts Now* slightly less likely still – indicating that those choosing to engage with the log did not seem strongly predisposed to reflection – or they simply preferred not to record their reflection. *Future Considerations* were entered about a fifth of the time, demonstrating little forward planning (notable when considered with the PDP results below). Finally *Contributing Factors* were only described about 5% of the time, perhaps revealing trainees did not feel able to, or did not value, describing the event within a wider context.

The data showed some similarities between the population and what was reported in

the wider literature; however, the engagement of the trainees and supervisors with the Foundation Programme and/or ePortfolio was often very limited and therefore close correlations with the literature (Edwards et al., 2003, Ehrlinger et al., 2003, Lane and Gottlieb 2004, Langendyk 2006, Ehrlinger et al., 2008) could not be identified. It should also be again noted that the literature typically described succinct studies of self-assessment in an educational intervention, rather than an examination of an entire year's training data.

6.3.2 PDP: Does (Self-Assessment) Promote an Appropriate Change in Learner Learning Activity?

The Personal Development Plan (PDP) delivered the facility for trainees to set out what they thought their learning needs were and note how and when they might be met. There was a "mandatory" section, but the Foundation Programme required only "evidence of use throughout the year". This ambiguity is unlikely to have motivated the actual use of the item, which was sparse.

Just under a third of both high and low self-assessor groups entered any PDP items, with number of total entries being slightly higher amongst low raters. Whilst this could be viewed as demonstrating more engagement, it is difficult to assert this objectively given the low level of engagement. A corruption of the PDP data stored also meant entries were only saved until mid-way through the second post. PDP items were also set to be tagged as Open (default), Completed and then Closed. A similar number of groups B and C (44%, 42%) returned to entries to mark them as completed (note this was for the entire year as the corruption did not affect entries already in the system) and 21 and 24% as closed. The fact that under half of the trainees (whether high or low self-assessors) completed their records draws attention to the fact that only a minority of the minority that engaged with the PDP saw it through to the training year's end.

Although both high and low self-assessors engaged more with the PDP than the population as a whole, the infrequent and erratic use of the tool made drawing any further conclusions about the groups impossible. In the examined year of data the PDP clearly did not elicit any substantive engagement from the majority of trainees who did

not use it to plan and link with their wider assessment and learning, and no substantive conclusions could be drawn about the identification of learning needs of different groups of self-assessors using the Foundation PDP due to the low levels of engagement with the tool, and to a smaller degree the incomplete data set. The infrequent use of the PDP could be attributed to many factors. These include the importance it was viewed with during the 2007-08 training year. This varied widely, as the whole scale adoption for Foundation was not uniform, even in its third year. Differing local conditions could be readily exacerbated by differing opinions of the value of new programme, especially within the priorities of busy hospital wards. The placement of the PDP itself within the ePortfolio was designed in 2007-08 to be less prominent then other components (such as assessments), which certainly would not encourage its uptake.) Finally, the fact that it was optional at the time would not encourage full engagement.

The engagement with the PDP, as a measure of self-assessment promoting appropriate change in learning activity, was sporadic. Some tentative connections with the systematic review's findings could be drawn, but for a variety of reasons lack of engagement prevented a comprehensive comparison.

6.3.3 Educational Supervisor Report: Does (Self-Assessment) Improve Clinical Practice?

Supervisors' Reports were examined to determine if self-assessment made any difference to improvements in clinical practice with their content being as close to an objective measure of clinical practice as possible (a *de facto* "gold standard"). When low self-assessors were confronted with evidence (in two posts) that they were rated higher than average by their supervisors, it was notable that half did not recalibrate themselves as higher in their consequent self-assessment(s). This is a notable difference from the literature, in which they did this in greater numbers. Less surprising were the high self-assessors, of whom half still rated themselves as high despite being confronted with opposing external assessment evidence. It would be expected, from the consensus of published studies, that only those that improved their base skills would be able to more accurately place themselves.

Supervisors could comment on every report (unless the score was <4 in which case they had to enter a comment, again perhaps unofficially contributing to the setting a score by which some supervisors would not rate less than), but it was not uncommon for nothing beyond a score to be entered in this data. Comments on the high self-raters were complex, with many caveats, mention of gradual improvement and learning and criticism balanced with potential for improvement. Conversely, low self-raters who had high Supervisor ratings enjoyed extremely positive comment, praising their skills in relation to their peers and in one case citing self-belief as being the only thing they needed to work on. The systematic review found good evidence that regular structured feedback was key to improving self-assessment, but it was clear that in this year of Foundation training this was the exception rather than the rule.

Although there was little recalibration within the chosen training year among the high and low quartiles of self-assessment the comments match Groups F and G as expected and it could be argued that given Foundation was relatively newly introduced. Adherence to process and ePortfolio usage was sporadic, it is understandable that predicted self-assessment patterns could not be fully observed, given widespread local variation in support and practice.

6.4 Foundation ePortfolio as medium for self-assessment

Like self-assessment, the use of portfolios has been strongly advocated across the health professions despite the lack of a comprehensive examination of the evidence for their effectiveness. Within a professional context the use of both self-assessment and portfolios might not be welcomed by all, but even amongst the majority of the sceptical they have been accepted as inevitable tools. But with this acceptance there is too often the problematic assumption that the professional can both use a portfolio, as well as self-assess, proficiently. The case study provided an opportunity to do what the systematic review noted was largely absent from the published literature, namely an objective examination of portfolios in practice.

Portfolios have in the past been used as places of storage – paper or electronic filing

systems. In the last number of years however the rapid substantial growth of more sophisticated e-portfolios has opened many possibilities (structured assessments, reflection, learning and professional planning) that the paper medium, or previous simple e-systems, inhibited or made impossible. The portfolio should perhaps be seen as a 'tool' to support education, not an educational instrument in itself.

Many factors have contributed to the huge expansion of e-portfolios (Gray, 2011). Disparate localised paper systems could suddenly be easily standardised for groups (e.g. Foundation trainees) in the now ubiquitous web browser. From desktops to laptops to smartphones, access to the web via a browser has become accepted as an essential means of communication, and as connectivity continues to improve so does access to one's e-portfolio at the point of practice – but also when the individual has the time and place to reflect.

The amalgamation of data in a single place that could be readily and/or automatically interrogated meant that poor performance or detailed data analysis became possible in a way that was never previously possible. An e-portfolio can easily be enabled to provide a flagging system that instantly contacts relevant supervisors by email or SMS when a poor score is registered against a trainee/student. Similarly, quality reports can be run for any defined group as the data held by an e-portfolio system can be queried as a regular or one off report.

Trainees across the health professions are increasingly expected to maintain portfolios for specified periods of training to collect and collate assessments (including self), as well as learning, appraisal, and annual review. In many cases this is now a formalised requirement e.g. A Guide for Postgraduate Specialty Training in the UK, 2008, "The Gold Guide". Medical, dental and other professional trainees are expected to regularly present their e-portfolios for review to supervisors, demonstrate progress for ARCP (Annual Review of Competence and Progress) and similar procedures, as well as having them used for sign off of satisfactory completion and registration with regulatory bodies (such as the GMC). As noted in Chapter 3's review of the evidence, mandated use will obviously increase uptake, but does not ensure engagement with anything past the required minimum. The case study showed this, with a majority of trainees

doing little more with their ePortfolio than was absolutely required.

Clinical exposure within training is under continuous pressure from financial constraints and the implementation of legislation such as the European Working Time Directive (Department of Trade and Industry, 2003), all increasing tension around reduced resource. There is a growing amount of evidence, much of it still anecdotal and not formally quantified, that when optimised electronic processes are far less time and resource intensive than manual ones. E-portfolios have been identified as tools to assist learning, appraisal and assessment during training, but also as potential tools and vehicles for such things as revalidation. The expectation that qualified (and to some extent qualifying) individuals should be able to assess themselves within electronic systems has come to be accepted (albeit sometimes reluctantly) by primary users, educators and the regulatory bodies.

In addition, self-assessment has come to be seen as central to lifelong learning in the health professions (Duffy and Holmboe, 2006), and its increasing appearance within electronic portfolios provides opportunities that are inherent within this flexible medium. Like the previous BEME review on self-assessment by the quality and heterogeneity of the papers that matched the review's questions were problematic. A meta-analysis of data was therefore impossible, but conclusions from the evidence could be drawn from a mixture of critical analysis of quality and holistic relevance.

A well organised implementation was seen as critical for the uptake of portfolios, particularly with mentors or supervisors who are willing to engage with feedback and other interactive processes – arguably, in the case of Scottish Foundation this was sporadic at best. There is some evidence that users feel more responsible for their learning with portfolios, and can be simultaneously sceptical and appreciative of them. Again, this requires engagement with the uses, not least to garner their opinions if not full fostering of learning, but with most systems being imposed from above and has been left largely unexamined (arguably until comparatively recently). The widespread geographical differences in uptake and practice discovered in the previous chapter continue to be noted across UK Foundation today. As Foundation was new and adapting as it was being implemented in high-stakes clinical environments, it is

surprising that its usage varied widely depending on local conditions.

Anecdotally, users of the Foundation ePortfolio reported similar things to what the review found with electronic portfolios being viewed as more flexible than paper, are used longer, and are seen to be better for feedback and reflection (van Wesel, 2008; Antonelli 1997; Sweat-Guy, et al., 2007: Driessen et al,. 2007b). Actual assessments scores were well correlated for both media, but there was no formal comparison between the brief use of paper versions in the two preceding years and the first universal use of ePortfolio during this training year.

Amongst the gaps in the evidence this review identified are the need for generalisable evidence over longer terms and the genuine outcomes of portfolio use. The use of a full year's e-portfolio data from Foundation medicine provided an opportunity for analysis of both self-assessment and portfolio use in a natural research laboratory where activities could be monitored in relation to each other.

6.5 Summative Assessment

Summative assessment is a controlled, standardised and traditional method of judging learners. The process is frequently high stakes and for both the assessees and regulators, particularly in the medical and dental professions. It has its limitations however. Whilst its role in accreditation is key, in isolation it can be seen to inhibit broader, or lifelong, learning in that learners will alter their behaviour to focus on the particular tasks they will be judged upon. The vital role of feedback in summative assessment mentioned in the literature (Antonelli 1997, Cox 2007, Lockyer 2005) is well supported by an e-portfolio, as the platform can be configured to imbed or even require feedback within prescribed intervals of an educational programme. A proficient system will facilitate further opportunities to mentor the trainee, returning to the assessment at future dates. In the training year examined the intermittent use of feedback exposed practice that was far from the intended ideal.

The NES ePortfolio is trusted to facilitate, record and collate hundreds of thousands of summative assessments annually. The continuing expansion of the ePortfolio is testament to the fact that it can effectively deliver the tools for summative assessment

and enable the monitoring and comparison of tens of thousands of individuals on a secure and standardised platform. However, sheer volume does not indicate that it is being used effectively.

Chapter 3 illustrated that reliability of summative assessments within e-portfolios varies widely, and the examination of the usage and scoring Foundation data would echo that concern. Enabling self-assessment technologically (even on a very large scale) may in fact be easier than ensuring that self-assessment is used in a consistent and educationally valid way.

The literature strongly recommends increasing reliability by having multiple raters and triangulation with other assessments. In the training year examined the MSF tool did indeed have multiple raters and a supervisor could readily compare MSF with other assessment results. But in practice questions can be raised about the tiny demarcations between MSF scores, the lack of assessor training and the fact different professional groups rate in different ways (Whitehouse *et al*, 2009). Similarly, the data showed many Supervisors only engaged with the educational processes at the bare minimum required. Therefore it cannot be assumed they compared results between assessments, and much more could be specified for e-portfolios to better facilitate the comparison of data for individuals and groups.

Beyond the facilitation and administration of self-assessment, an e-portfolio can offer more. The electronic platform also enables the association, or linking, of summative assessments with other portfolio components, making the individual's summative assessments more than a list of results, but part of an integrated learning record over a longer period of time. Whilst improvements have been made to the system since 2007-08, time and resource continue to prevent e-portfolios from achieving their full potential.

6.6 REFLECTION

Reflection is seen as key to experiential learning (Maudsley 2000; Sobral 2000), and there is an increasing amount of evidence that reflection can help students learn from their clinical and non-clinical encounters. E-portfolios aid and evidence reflection

providing the learner a structured environment to compile and associate the assessments, feedback and encounters they have learned from, with each other or against generic competence frameworks or curricula (Driessen 2008). An e-portfolio also allows the learner, as well as their supervisors, to plan and monitor future goals. Structured reflection has been widely seen to benefit the creation, maintenance and achievement of learning objectives (Norman, 2004).

Within an e-portfolio, learners can recognise opportunities for reflection themselves or be reminded to at prescribed intervals. The ability to record a self-assessment when an individual desires, or their responsible supervisor wishes, gives the assessment an immediacy that cannot readily be replicated in another medium, offering the learner a significant degree of control. In practise however, the 2007-08 training year saw very infrequent use of more than the minimum required number of self-assessments and no strong evidence of significant use of the entire ePortfolio as a reflective tool.

It could be argued that the often brief comments entered in the self-assessment forms would have been greatly enhanced if the design of the Foundation portfolio supported reflection. For example, trainees expressing self-doubt could link their areas of concern to other assessments, prioritise them in a PDP, be encouraged to expand and record their reflection, and even tie in with relevant educational opportunities, such as upcoming lectures or e-learning content.

They enable the expression of a wide range of personalised experience in conjunction with linking to standards and curricula. By fully utilising the electronic media, the curriculum, which could often sit neglected in paper format, can be fully integrated with assessment, reflection and learning. But the electronic medium is by no means a simple answer; for example, documented feedback is of little use without evaluation and active reflection to alter practice. An e-portfolio needs to enable all those involved in the educational process, rather than become an artificial environment imposed upon learning. Careful consideration should be given to pedagogical design to avoid merely creating a “tick-box” application, but a web-based format can integrate diverse separate components and add value to educational processes.

Reflection requires the identification of learning needs and the ability to assess one’s

own skills. Neither of these things comes readily to most individuals. An e-portfolio can however facilitate both, by (for example) embedding prescribed guidance and feedback from mentor and peers in relevant sections and all could potentially benefit from a platform that integrates and facilitates what we know aids the reflective process. This could include a variety of activities, such as identifying educational gaps and planning education to remedy them. In Foundation, this was what was intended of the PDP, but as an *optional* semi-integrated and somewhat side-lined section, it never lived up to its potential in 2007-08. Consequent annual changes have improved this, and the uptake in its use has been considerable. More can always be done, however, with links between contents, results and individuals within the system, but as always this is dependent upon timely analysis of usage, piloting of refinements and collaboration between stakeholders and developers.

6.7 Working Environment

Workplace-based assessment continues to be seen as critical to the education of healthcare trainees, and e-portfolios are uniquely placed to support them. They can be initiated by trainee or trainer and the e-portfolio can support and structure initial and consequent feedback. Although only part of an overall assessment system, via an electronic platform WBPAs can quickly identify poorly performing trainees so the relevant people are informed and support can be offered to improve their work.

Self-assessments can obviously also be WBPAs. Ideally they will trigger other learning events, including consequent WPBAs. Integrated within an e-portfolio, the areas for improvement identified by WPBAs can be linked to (for example) Professional Development Plans to inspire a trainee to improve; similarly they can be linked to other components such as a curriculum or reflective logs. Again, in line with the literature, Chapters Three (Ryland et. el., 2006; Snadden & Thomas 1998; Murray 2007) and Four both demonstrated that mentors or supervisors need to be fully engaged with the processes, or actual practice will not change.

Mobile devices offer great potential to perform and record assessments at the point they occur. Whilst the complex realities of working on a busy ward will obviously often

preclude their immediate use, as will the physical and security confines of many hospitals (Vallis, 2008), hand-held devices such as smart phones and iPads are rapidly gaining an increasing share (approximately 16% of ePortfolio traffic in January 2014) of usage and the trend can only increase with the development of Apps and HTML5. Connectivity challenges are slowly improving across the NHS but the ability for a handheld device to record an assessment or reflection offline, only to sync with the live database when a secure connection is again established, increases the utility of these devices and enhances their component parts. The ability to readily conduct assessments at point of practice offers the key advantage of the assessor being able to conduct the assessment at the time of the event, rather than recollect and record their perceptions at a later time. This immediacy can be seen to improve the accuracy and value of the assessment and consequent feedback (Russell, *et al* 2006; Norcini and McKinley, 2007).

Assessments, self or other, are frequently done at point of care. Despite the obstacles inherent with mobile technologies in hospital settings, there is a growing expectation from users that wireless internet access should always be available, which would enhance the topical accuracy of assessment, provided the e-portfolio supported point of practice assessment. Currently, systems such as NES ePortfolio support checklist type assessments well, but more reflective or introspective assessments that require the entry of large amounts of text are impractical on hand-held devices. Applications such as the iPad that reside between the phone and the laptop/pc show great promise and are starting to achieve widespread use. Voice recognition software can also alleviate shortcomings of smaller devices, but these have yet to appear in any substantial way. A common scenario can be imagined in the near future where assessments are conducted at or immediately after a clinical encounter entirely by voice.

As noted in the portfolio systematic review, to be accepted across an organisation, an e-portfolio system (as well as the functions and processes it supports) requires buy-in across an organisation, particularly at the higher levels. Partial or incomplete support almost certainly means a slower and less successful uptake of a system, even if parts of

it are mandatory. The data from the early use of the NES ePortfolio demonstrates just this, with regional variation in completion rates for all component parts. Engaging with the stakeholders before a system, or changes to a system, are introduced is critical to realising its full potential, especially when competing for time in a clinical workplace.

6.8 ACCREDITATION

The need for quality assurance in healthcare education is paramount. Previously, paper-based systems have meant that any collation of data took weeks or months, whilst on an electronic system this is done automatically along pre-defined queries. Assessments for individual trainees that are of concern are instantly identified to relevant parties so assistance can be remedied in real time. Comparisons between groups or geographical areas can equally be done with minimal effort. Self-assessment could readily be enhanced by the considered exposure of the individual to wider assessment scores, such as the comparison to live national benchmarking.

An e-portfolio can provide a rich and representative picture of the learner for their transition through training or progression throughout their career. They are increasingly used by regulatory bodies to capture the information required for accreditation, as well as providing snapshots of group use (compliance with requirements, scores across specified sub populations, etc.). With self-assessment frequently forming a core part of accreditation, systems being designed for accreditation must take into consideration its strengths and weaknesses. The dangers of not properly considering both the process of self-assessment and the environment in which it exists, is that its use will not be educationally valid or useful and/or it will be seen as a tick-box exercise which would ultimately fail to safeguard standards and improve practice.

ePortfolio was being used in 2007-08 to certify satisfactory completion in Foundation medicine, but also more widely in medicine and dentistry as well. For a variety of reasons, some unknown, trainees achieved certification competence despite not having met the full requirements. Whilst these exceptions are far fewer today, it reinforces the point that an electronic record is a tool that requires change in actual

practice if it is to achieve its full potential.

Accreditation raises further issues as well. Individuals are far less likely to engage with non-mandatory items (such as reflection) if they do not trust the data holder or feel their data might be shared without their consent with regulators. Similarly, as e-portfolios are extended (and potentially linked) from university to training to practice, users may not wish current supervisors or the regulators to view items in their past. Data protection issues, as well as related issues such as data migration and archiving, must be resolved in a transparent way to encourage uptake of a system and the processes it supports, as many individuals are not comfortable with various bodies holding their data beyond a certain time, or making it accessible at a future date.

6.9 ENGAGEMENT

The completeness of the ePortfolio dataset for Foundation trainees in Scotland and its ability to store accurate and reliable data, according to observed outcomes for individuals, was tested during this study. The system itself was proven efficient, robust and fit for purpose, but this did not necessarily result in users' engagement with it. Chapter 3's review found similar issues around compliance with portfolios varying depending on whether requirements were mandatory (Pearson & Haywood, 2004; Snadden & Thomas, 1998; Smith & Tillema, 2001; Murray, 2007).

The 2007-08 training year was the second year all Scottish Foundation trainees were required to use the system, and the third year of the large scale changes brought by the implementation of the new Foundation Programme. This evaluation found there was a high level of participation with the system by the majority of trainees, including both mandatory and non-mandatory elements. Stored data (as entered by trainees) were found to be accurate and erroneous entries were (relatively easily) identified and resolved by local administrators. There were regional variation in the adoption of procedures, and these were often found to be due to competing local issues or temporary problems with IT limited access. As noted previously, the diversity of training environments found across Scotland means practice must adapt to local conditions – a large hospital in central Glasgow provides a far more diverse range of

posts to those available in remote and rural areas.

Since the year of study, engagement has increased and clear national guidelines on requirements for trainee sign-off have now been implemented. These have improved compliance and the quality of data recording. These conditions were not necessarily synchronous with the early years of Foundation, and the discrepancies in ePortfolio usage and compliance amply illustrate this. Training provision on ePortfolio has also expanded and continues to drive improvements in use of the system; however, training to use software is comparatively

Mandatory requirements (assessments, meetings, declarations), especially when well organised, will improve levels of engagement and this is an obvious contributing factor to the training evaluated year and the consistently improving levels of engagement since. But the levels of use beyond mandatory requirements, as well as the use of optional items, was often high and has been consistently increasing over time. A longitudinal evaluation of engagement with self-assessment in the ePortfolio, and how this related to the other activities within the ePortfolio (PDP, Educational Log, Curricula, uploaded evidence), would be a valuable exercise to inform Foundation ePortfolio, and provide timely evidence to the wider medical education audience. Since the training year examined in this study, there has been considerable change to the content of the Foundation ePortfolio which is discussed below.

6.10 Learning Support

There is an increasing recognition that the process of learning is as important as the end result (GMC, 2012), and an e-portfolio is ideally situated to document and enhance the entire learning processes.

E-portfolios provide the flexibility that can enable the learner to readily share ideas and receive prompt feedback on them. Through e-portfolios, this could be extended beyond the individual to incorporate various sized groups, as specified by the stakeholders through the software. E-portfolios enable the learner to reflect on experience and plan in response to that experience. Provided the appropriate tools are available within the software, the learner can aggregate digital items in purposeful

ways to present to whatever audience is required.

If the full value of self-assessment is to be realised the individual needs to be able to take appropriate action based upon their results. If the self-assessment tool is integrated, or at least linked, to relevant related resources, the value of the electronic platform is far increased. For example, if a trainee is readily able to recognise and record weakness in a particular skill, the ability to then note it as an action point in a PDP or curriculum node, message their tutor to note the learning need, upload items associated with the need and automatically link to relevant e-learning material, then the e-portfolio truly becomes a supportive tool taking full advantage of its medium.

At an institutional or regulatory level, the data compiled by e-portfolios can provide significant longitudinal depth to improve and enhance education, allowing stakeholders to understand and plan on an evidence base that was not previously available. The UK Foundation ePortfolio, like most e-portfolios, undergoes iterative alterations at regular points to improve the product and best serve the changing educational environment. Contributions by portfolios to organisational practice were also noted by Cotterill et al. (2007) and Swallow et al., (2006) with regards to the planning and organisation of learning.

Significant hurdles remain to making e-portfolios reach their potential to fully support education. This is true both in terms of integrated functionality across educational activities, as well as reportage of the data e-portfolio systems themselves collect. Much of the work to be done depends on organisational will and availability of resource, as the technology can already support these improvements and continues to evolve to offer more flexibility and functionality.

6.11 E-PORTFOLIOS AND E-LEARNING

In the context of the assessment process, an e-portfolio is able to streamline evidence identification and validation, and enable assessors to effectively make judgments about the authenticity of evidence when it is verified through existing legitimised sources, such as Student Management Systems (SMS) or Learning Management Systems (LMS).

Many organisations promote the LMS as a single platform for teaching, learning and assessment, but there is a tension between the formal institution-centric role of LMS versus the learner-centric role of the e-portfolio. Traditionally, the LMS was seen as giving a degree of control to the organisation and stakeholders that an e-portfolio could not, but this could be viewed as misguided given portfolio content can be entirely prescriptive and controlled, though is usually a mixture of organisationally and individually generated.

Increasingly, the lines between e-portfolio and LMS are being blurred, as e-portfolio systems are delivering more and more learning content. Increasingly, e-portfolios are being seen as the central tool for the learner, collecting and collating reflection, summative assessment and personal development, whilst the LMS is being viewed as a storehouse for learning content. As more e-portfolio systems are able to manage and deliver learning, they will increasingly be seen as the medium for conducting learning as well as assessment.

The shortcomings of the LMS model are well documented (Emory 2007; Schroeder *et al* 2010). They are very often seen as stores of static content (e.g. PowerPoint presentations and lecture notes) and the amount of usage they generate is questionable. There is also the impression that an LMS is an inflexible tool that binds learners in prescribed paths. Institutions are confronted with supporting their own LMS whilst learners opt for external blogging and social networking sites top communicate with colleagues and create and share content. Critics of the LMS model argue that they in fact inhibit innovation as they cannot evolve as quickly as web-based technologies. To date, in the UK healthcare sector, assessments have been done by systems external to LMS.

By definition, an LMS is about an organisation's desire to manage learning, where more recent activity has stressed the emphasis to be placed on the individual, or the "personalised learning environment". The latter can be supported, or even entirely delivered by an e-portfolio system, with the educational emphasis shifting from a single organisation to the learner themselves.

A plethora of free online tools (dictionaries, thesauri, scientific calculators) already

exist, further weakening the argument that a centralised learning management system is necessary over individual collaborative tools in conjunction with the learners' record. Learners fully expect to switch between various web-based services and their own laptops can be viewed as de facto personalised learning environments.

The LMS will certainly continue to exist for some time, however, as institutions have invested heavily in them. They also still provide functionality, such as course administration and collection of fees that is not readily provided by personalised learning environments. Despite their shortcomings, it is most likely that e-portfolios will have to exist alongside, and exchange data with, the LMS for the foreseeable future. Organisations will continue to have to decide where assessments and reflections are most suited for delivery and completion, given that learning content is increasingly reached through multiple access points.

The 2007-08 training year was the final year that the Scottish Foundation ePortfolio was linked (shared common authentication) with the parallel LMS DOTS (Doctors Online Training System). As noted in the previous chapter, ePortfolio migrated to a new version in 2008 based on updated technologies, whilst DOTS remained written in older and less compatible technology. The decoupling was necessary for ePortfolio to meet the changing demands of the user base, but meant trainees, supervisors and administrators lost the benefits of a combined system. In August 2012 DOTS was shut down, with its functionality being transferred to a new area within ePortfolio: "Learn". This will be an integrated Personal Learning Environment within an e-portfolio, which will be closely monitored to determine if it should be extended beyond Scottish Foundation trainees. The Foundation Programme year examined in Chapter 4 recorded assessments and reflections as isolated events. Improvements in functionality and processes have seen the Foundation ePortfolio evolve into a system that links all components and has enjoyed steady increases in use year on year. "Learn" now integrates course modules within the ePortfolio and provides an integrated user experience and the chance to directly access learning relevant to ePortfolio items. As a learning environment, ePortfolio could come to accept increasing amounts of external content which would allow supervisors and peers to see not only the result of the

learning activity, but also the process in which the knowledge and skills were acquired.

6.12 Web 2.0 / Social Media

Technology has been rapidly altering the way people communicate. Social networking technologies (often described as, or affiliated with the term Web 2.0) have proliferated over the last number of years, and the ability to create and share experience in these various mediums has become expected.

Web 2.0 technologies, or social software, are defined as “web applications that facilitate participatory information sharing, interoperability, user-centred design, and collaboration on the World Wide Web. A Web 2.0 site allows users to interact and collaborate with each other in a social media dialogue as creators (prosumers) of user-generated content in a virtual community, in contrast to websites where users (consumers) are limited to the passive viewing of content that was created for them”. (Wikipedia, 2012)

In many ways e-portfolios are Web 2.0 technologies, in that they allow the user to assemble, organise and present learning that has occurred out of formal educational settings. These technologies often cite the benefits of immediacy, informality and access to emotional support as being key to their support of collaborative learning, as well as self-reflection. In many ways the ubiquitous Facebook is in fact a type of e-portfolio, though the informal and near purely social use of this tool sharply distinguish it from professional e-portfolios.

Other Web 2.0 tools (wikis, blogs and social networking) create previously impossible chances for creating and exchanging content with any participating group or individual. Trainees now arrive in the health service with proficiency in these technologies, as well as the expectation that the technologies will be available to enhance their work experience.

Social media is now being extensively used within the current ePortfolio, with tools such as Twitter enabling trainees to feedback suggestions to both educational stakeholders and the technical team. This has provided an opportunity within what has been seen as an often anonymous system for an individual to contribute and feel they

have a voice. Increasing individuals' ability to contribute is enabling them to feel their involvement is more personal and immediate, which in turn should have benefits in encouraging greater engagement with the processes.

Similarly, new technologies such as Mozilla's Open Badges present e-portfolios to accept and share formalised and agreed learning achievement from other systems, and in turn would give an individual or tutor much more recognised learning content to meet gaps identified by self or other assessments.

The collaborative nature of Web 2.0 technologies can be seen to promote self-assessment in that inhibitions about criticising one's self can be offset if the individual is part of a wider community (such as a blog) where the practice is shared. Although informal, the juxtaposition of quasi self-assessment/reflection components with other social media could be argued to provide a form of triangulation that would enhance the value of the constituent parts. In practice however, it is likely that trainees will continue to want to keep separate their professional and personal online identities. E-portfolios are well-suited to support user-generated content and can easily provide and enhance personal reflection. However, whilst some of the functionality of e-portfolios and sites such as Facebook might be similar, and the reproduction of Facebook features would be comparatively simple, there has never been significant demand for the replication of this in the NHS ePortfolio from the user feedback exercises. This does suggest that trainees see ePortfolio as their professional representation of self, which they want to be separate from their personal lives within a wider social media.

6.13 Technical Dimensions

Within healthcare too much data are generated and distributed in multiple disparate repositories – e-portfolios being just one constituent type. Comparatively recently the concept of "Big Data" has emerged which aims to manage and harness the vast and rapidly growing amount of data being generated across the world in incompatible formats. This is omnipresent across healthcare and also applies to the large variety of disparate data collected, but not connected, within healthcare education.

The technology exists to source and compile data in real time, but there no history of interoperability between IT systems supporting healthcare education, and data are frequently entered more than once in different systems. A trainee's educational data will be found in deanery/College databases, e-portfolios, e-learning platforms, assessment systems, quality management systems, etc. frequently in multiple instances of the same type. Whilst the technology is in place for seamless data exchange, the required processes very rarely exist. What are required are data-sharing agreements, mutually agreed rules for custodianship of data and system tools for authentication, record matching, permissions, transaction tracking and audit.

Migration to cloud or virtual servers from the traditional static hosting arrangements will also become the norm in the near future. The enhanced processor power cloud and virtual hosting arrangements provide will provide flexibility for all users, with the ability to create near immediate analysis of comparative data within the vast tables stored by the system. Parameters for appropriate release of data would have to be set, but the ability to triangulate self-assessments with other scores from other systems could provide the desired reference points to make self-assessment more accurate and meaningful.

6.14 LIFE-LONG LEARNING

There is general consensus that the healthcare professions need lifelong assessment, as trainee assessment and all that it entails does not last the 35-40 years of a career (Miller, 2005; Shaugnessy, 1999; GMC 2009). Self-regulation (based on self-assessment) is therefore seen as a pillar of lifelong learning.

Given their flexibility and portability, e-portfolios are a natural tool for life-long learning. They can easily facilitate information sharing for interviews (for such processes as the ARCP, Annual Review of Competence Progression) as well as appraisal. As a learner's repository, an e-portfolio can be aggregated to provide evidence for multiple audiences including educational supervisors, tutors, peers and employers. Improved self-assessment, ideally triangulated with other feedback, would be of great benefit to the learner through the continuum of their career.

As there is no single e-portfolio system that could support all possible needs, nor is there likely to be one, the viability of life-long learning rests upon the development of data exchange between e-portfolio systems. At present there is no recognised specification, much less international standard, for the exchange of data between e-portfolios. Attempts to date have been academic exercises that have yet to have any significant adoption by stakeholders (ref trajectory). There is currently a consolidated attempt between the Royal Colleges of Medicine, the NHS and private companies to agree a standard for medical e-portfolios to exchange data, but this is still under development and would require the buy-in of all interested parties if it was to truly facilitate the learner's life-long journey.

It should be noted however that when discussed with providers such as NES ePortfolio, organisations that require e-portfolios, as well as their users, consistently say they do not want a blurring between their professional e-portfolio and the social media profiles they create. Whilst the ability to include relevant educational experience and achievement from respected sources is highly desired, a student or trainee does not necessarily want a supervisor to link into their personal Facebook content.

Personal development can be greatly enhanced by a system that supports reflective practice, collaboration with peers, and the organisation and presentation of achievement for daily use and subsequently reliable recall over indefinite periods of time.

Critically, future developments in e-portfolios need to imbed mobile technologies to make the lifelong record omnipresent. A portfolio that is ubiquitous, and would therefore support learning whenever and wherever it occurs, is key to ensuring the individual learner embraces it for the long term.

6.15 Performance Management

E-portfolios enable users to document and share their achievements within the context of their experience whilst providing the opportunity to reflect on their experience connected to the wider learning environment. Data entered into assessments tends to be a standard of what's achieved whenever a supervisor is

willing to take an assessment. This may mean that it documents competence which is already achieved, rather than the learning curve. During the period of study anecdotally, there was widespread reports that work-placed based assessments were repeated until a trainee reaches a degree of competence at which point they and their supervisor feel they could enter the result into the ePortfolio.

Ideally, an e-portfolio system would be optimised (through flagging of incidents of concern, scores that fall below agreed thresholds) to enable the rapid identification of those most at risk of failure or incompetence. Many electronic systems of assessment, UK Foundation amongst them, have extensive and elaborate tools and processes that consume large amounts of time and resource – especially when there are tens of thousands of users. With the technology already in place to help determine the number of assessments actually required to assure quality, the people and organisations involved need to use the technology to its full effect. To some extent this is already happening, with deaneries often doing scheduled informal trawls of the data. But this falls short of having the system automated to scan for problems and inform the relevant parties when required. Issues could also be automatically passed on to other parallel processes, such as the ARCP.

Ideal data sets to manage performance would be: personal data (in a single source), learning experience (e.g. experience with patients), technical procedures (e.g. logbook type entries), learning achievement (certified content), assessments (self and other), and reflection (structured and unstructured). It would be a mechanism for real time data collection and/or particular collections of data specific for the purpose of competence assessment.

Competence has been described as a constituent part of lifelong learning, rather than the achievement of a prescribed state (Leach, JAMA 2002). To improve, the individual needs to gain self-insight and the ability to adapt to the evolving work environment. An e-portfolio could be viewed as a tool for educational diagnosis: an assembly of a myriad pieces of evidence for the learner to evaluate themselves over a continuum.

A trainee's assessment as being competent does not extend throughout the duration of their consequent careers. Self-regulation, including self-assessment, is an important

if not crucial part of maintaining competence. An e-portfolio can collect and collate data required for multiple purposes. Often the e-portfolio is seen as a high stakes and summative tool, testing for minimal competence and knowledge as a standardised measure for institutions or regulators (a specific measure of competence). But an e-portfolio can also deliver for the individual trying to gauge, maintain and improve one's own skills and behaviour (an evolving process).

Healthcare professionals generate large volumes of data throughout their daily practice. As they move through their careers, educationally relevant data will inevitably reside in many separate locations. An e-portfolio system does not need to replace these repositories, but ideally it will retrieve relevant information and collate it in real time.

Interoperability between electronic systems can be problematic and time consuming on various levels; however, it can be approached incrementally between receptive organisations and systems. Exchanges, as described above, would involve agreements to share data, either for particular instances or a total exchange. The specifications for this would need agreed, tested and maintained – potentially by an external entity that acted as a conduit for the collaborating partners. The specifications would need to exactly map identified fields, ensure secure authentication, fully define business rulers, agree access rights/permissions, record all transactions and additionally provide audit functionality. Although the work is hardly insignificant, the advantages of having education as a career-long continuum, rather than isolated events or periods is considerable as individuals would have a lifelong record and organisations a rich set of longitudinal data for planning and quality assurance. The e-portfolio would be the bridge across the continuum, helping education being seen as a continually evolving process across one's career, whilst simultaneously giving immediate access to relevant data held within the system.

The 2008 Tooke Report highlighted a number of issues around assessment methods, as well as the use of portfolios, in UK Foundation. These included, "a lack of clarity about how the portfolio is to be used and how it can be assessed", multi-source feedback needing to be "as comprehensive as possible" and "lack of assessment at the cognitive

level". The evidence base around portfolios demonstrates lucidity and precision in instruction and implementation are required for optimal adoption; similarly, there is growing evidence around the utility and effectiveness of self-assessment both in itself and in relation to other assessment tools, provided its use is informed and judicious. Given the money and time involved in formative and summative assessment in postgraduate healthcare education, as well as the high stakes nature of accreditation and recertification, it is imperative that all stakeholders make best use of the available evidence base to inform the creation and maintenance of both tools and systems.

7 CONCLUSIONS

7.1 WHAT CAN BE LEARNED FROM THE CASE STUDY DATA?

This extensive evaluation of a year of Foundation Programme data presented some results which agreed with the self-assessment literature while others contradicted published findings. The fact that the data did not fully replicate the literature could be put down to several factors, including:

- lack of engagement with the new assessment processes of Foundation active at that point of time
- unfamiliarity with the new portfolio and/or electronic format (echoing the findings of the second review)
- predisposition of raters to define scores within a narrow "acceptable" range
- and the widely held belief that assessment scores should not be recorded in the ePortfolio until ratings within this band were reached.

Regardless of the reasons, it is clear that both self-assessment and the medium of ePortfolio were not being used to their full potential by the examined trainees. The first review highlighted (Section 2.8.8) that no papers focused on user perceptions of self-assessment and its acceptability as an educational tool. This was certainly the case in this Foundation year as no training or guidance in the practice was given to assessor or assessee. Its use and value was simply assumed, and how the evidence base on self-assessment was considered by those designing the assessment system, was not made available. Its place within the ePortfolio had similar echoes with the second review, in that the key factors in proper use and uptake (implementation, organisational support, time to complete, mentor/supervisor engagement) were variable.

The assessment system (self and otherwise, as well as related activities) delivered by a web-based portfolio did not reach its full potential on the ground. Nevertheless, this work did have correlations with the wider literature and raises many key points for consideration.

Despite the narrowly demarcated scores in MSFs generally, there was tentative evidence that self-assessment of clinical skills in these medical trainees' ePortfolio

replicated patterns in other studies. One area of notable variation from the literature was low early self-assessors strong predisposition towards practical skills as a preferred area of learning activity. This could indicate a predilection for experience over theory or other learning types; it might also support the notion that tangible skills were more readily self-assessed (Antonelli, 1997; Leopald et al., 2005; Weiss et al., 2005; etc.) and this group, having less confidence than their peers, felt the need to overtly demonstrate improvement.

Importantly, this work also revealed fairly widespread disengagement from the non-mandatory aspects of Foundation during the year under study. Although the mechanisms were in place to gather extensive data, limited use meant it was not possible to fully examine the unanswered self-assessment review's questions. This anecdotally supports the portfolio review's findings (Webb 2006; Snadden & Thomas 1998; Murray 2007) that implementation needs to be fully planned and supported by senior stakeholders to be truly effective.

This examination of the assessment scoring indicated that an overall improvement in the trainees' progression in competence was not detectable in the aggregate data. Direct comparison of this first and second year data however was not thought to be valid, as the competences assessed are different (as listed in Table 8), and in this study the available data comprised two distinct populations. Longitudinal, linked data analysis however would allow greater exploration of this area – particularly now there is more standardisation of assessments and a larger, more stable and complete dataset collected in ePortfolio.

As mentioned previously, a bias inherent in self-assessment is not necessarily the problem per se, so long as the bias is recognised and attempts are taken to counter it. As previously discussed in the literature and the case-study, self-assessment is typically introduced without widespread consultation or training. Educators can focus on developing the skills themselves, rather than an ability to self-assess; opportunities can be created to recognise the boundaries of one's own knowledge, so learners could induce the failures that they could learn by; and the focus on the accuracy of assessment should be on externals rather than the self. Additionally, other established

key factors to improve self-assessment, such as properly implemented feedback (Antonelli 1997) can improve the accuracy of self-assessment, but during this training year there was no agreed or planned support for self-assessment and its implementation and use within the year's assessments was extremely variable.

This study relied on a couple assumptions that could be challenged, such as the validity and reliability of the MSF tool used in Scotland at the time, as well as the use of supervisors' and feedback as the measure of actual clinical performance. As noted in the previous chapter, the Scottish MSF tool has since been replaced with the more widely used and tested TAB MSF tool, but for the year examined it was the only option for examining self-assessment in practice, as well as comparing it in parallel with external scorers. With hindsight, there would have been many ways to have had self-assessment contribute to trainees' education and progression in a far more consistent and meaningful way.

7.2 Can self-assessment be effective within an e-portfolio?

The case study starkly illustrated the difference between an intended education and training programme and what actually transpires in practice. But it is critical that the potential is not dismissed and taken as an opportunity to learn from the experience and make improvements for the future. This thesis does not make bold claims based on unqualified conclusions to the research questions it examined; however, the evidence examined did demonstrate that there was considerable potential for self-assessment enabled within an e-portfolio, but this potential has yet to be adequately realised.

Attitudes to both self-assessment (Eva and Regehr 2005, Section 2.8.8) and portfolios (Cross & White, 2004; Maidment et al., 2006) can be complex and are frequently not considered. Compliance with requirements would almost certainly improve if users were heard and understood. Similarly, engagement with self-assessment within the whole educational environment could improve if users fully understood what was required of them and why it is, and how they could benefit by understanding self-

assessment in context, rather than having it dismissed as another mandatory "tick-box" exercise.

The portfolio review found good evidence that the implementation method was key (Webb et al., 2006; Snadden & Thomas (1998); Kjaer et al. (2006)) to their uptake and proper use. Many studies reported that this was often neglected, and the anecdotal reports of local implementation of the ePortfolio in Foundation were the same, missing the opportunity to have the ePortfolio component assessments work to their true potential.

Related to this is the engagement of the supervisors and mentors. For a portfolio to be its most effective supervisors need to be actively aware and involved in the processes (Dreissen et al., 2007b; Webb et al., 2006; Ryland, 2006). Likewise, self-assessment improves with the full engagement of others (supervisor, peer) in instruction (Leopald et. al., 2005) and feedback (Weiss et al., 2005; Leopald et al., 2005; Antonelli, 1997). There is further (qualified) evidence that the use of video in self-assessment feedback can have a positive effect (Ward et al., 2003; Martin et al., 1998; Lane & Gottlieb, 2006). Provided the impediments inherent in many clinical settings (patient consent etc.) are overcome, the use of video (now near ubiquitous via smart phones) uploaded to a portfolio could be a significant step to making self-assessment, as well as other educational processes, more effective. Within the case study, training for supervisors and trainees could have been provided, with it highlighting the need for regular detailed feedback and the identification of skills than are better self-assessed (i.e. "practical / demonstrable over "soft"). Similarly, the pressure to inflate self-scores could have been offset by coordinated explanation of the benefits, which could also have offset assumed acceptance of self-assessment within the training year. And critically there was very little indication that self-assessment was being used as anything other than an isolated mandated task.

Peer assessment was revealed to be more effective than self-assessment (Rudy et al., 2001; Sullivan et al., 1999) in the first review, yet there is no reported use of both alongside each other to improve the accuracy of ratings within portfolios. Indeed both

reviews (Rees, 2005; Melville et al., 2004; Jarvis et al., 2004, Maidment et al., 2006) had high quality papers stating that assessment data should be triangulated to improve accuracy and reliability. An e-portfolio is an ideal medium to do just that, in the data can be compared nearly instantaneously with minimal technical development. Beyond data triangulation, additional functionality could be employed (such as linking self-assessment to other events, in particular professional development plans and reflection, and opportunities to improve the skills themselves) to improve the accuracy and effectiveness of self-assessment within a wider educational milieu. Again, in practice this is not currently happening in any significant manner but the ability to make self-assessment more effective and meaningful in a holistic context is readily available.

The amount of time available to complete a portfolio as intended was well cited within the portfolio review (Keim et al., 2001; Jensen & Saylor, 1994; Dagley & Berrington, 2005; Duque et al., 2006; etc.) and was broadly apparent within the data of the case study. Patient care will always come first, and education needs to be designed with a realistic acknowledgement of the pressures of the clinical environment. Assessment, self or otherwise must fit within this. Technological improvements such as handhelds and HTML applications enabling offline work and syncing to record once within a wifi signal, are helping, but so is the ongoing evolution of e-portfolio supporting assessment programmes.

7.3 FOUNDATION EPORTFOLIO 2005-12

The initial objective of the NES ePortfolio was to support effective summative assessment of trainees to demonstrate competence for satisfactory completion – and although the initial pilot did accomplish this, it did very little more than this. The primary purpose however did not preclude the development of other facets of what has become the NES ePortfolio: reflective practice, professional organisation and presentation, and (arguably) learning itself. An e-portfolio can focus on a single dimension, but is more often a compilation of multiple purposes. Whilst each function or process of the portfolio may be useful in itself, the true potential of the portfolio is

arguably not reached until its component parts are combined in meaningful ways. An electronic portfolio particularly lends itself to this, as data and formulae from multiple tables can be instantly combined for individuals or composite groups.

The growth of e-portfolios (NES ePortfolio is the largest, but there are many more) in recent years has been substantial. There have been a variety of reasons for this, including pedagogical change, technological opportunity, demands for quality assurance and the migration of learners between institutions and places or employment. Learning is increasingly viewed as central to the individual, rather than the traditional approach of collective groups. Learners are now expected to work with peers, reflect on their learning, and connect their wide-ranging experiences within agreed criteria. An e-portfolio is precisely designed to facilitate these activities, and integrating self-assessment is obviously an expectation of many systems.

The growth of e-portfolios can also be attributed to the expectation that individuals' records of learning and achievement should follow them through their careers. An e-portfolio enables one to collate evidence of accomplishments, as well as commenting on one's own attainment and presenting the evidence to relevant groups.

There have been significant changes to Foundation training and the ePortfolio since the training year examined, including two rewrites of the curriculum and adjustments and harmonisation of assessments. There are still some small regional variations in practice but core content and procedures are now uniform across the UK.

The ePortfolio now enables significant linking of evidence between sections, for example a trainee is able to upload files or associate forms with any node on the curriculum. The PDP is now, for example, much more central and relevant to the trainee and supervisor, rather than an isolated and often neglected feature. The platform also facilitates communication more readily between trainees and supervisors. These changes have seen a dramatic increase in engagement and use of the ePortfolio and an analysis of a more recent year would certainly provide more a comprehensive comparison.

As emphasised earlier, the ability to self-assess is not transferable skill nor one defined independently of the activity being assessed, but if the activity of self-assessment is

expected and routine (and facilitated through standard electronic environments) there is far greater potential to improve and learn from the practice.

In August 2010 many aspects of the Foundation ePortfolio were substantially altered by the UKFPO. These included the lessening of the role of certain assessment tools (e.g. DOPS), the addition of a clinical skills log and the redesign of the ePortfolio making the curriculum and PDP central to the process flow. These changes were implemented to engage the trainee (and supervisor) more closely with the competencies in the curriculum, and better integrate the many aspects that comprise Foundation training.

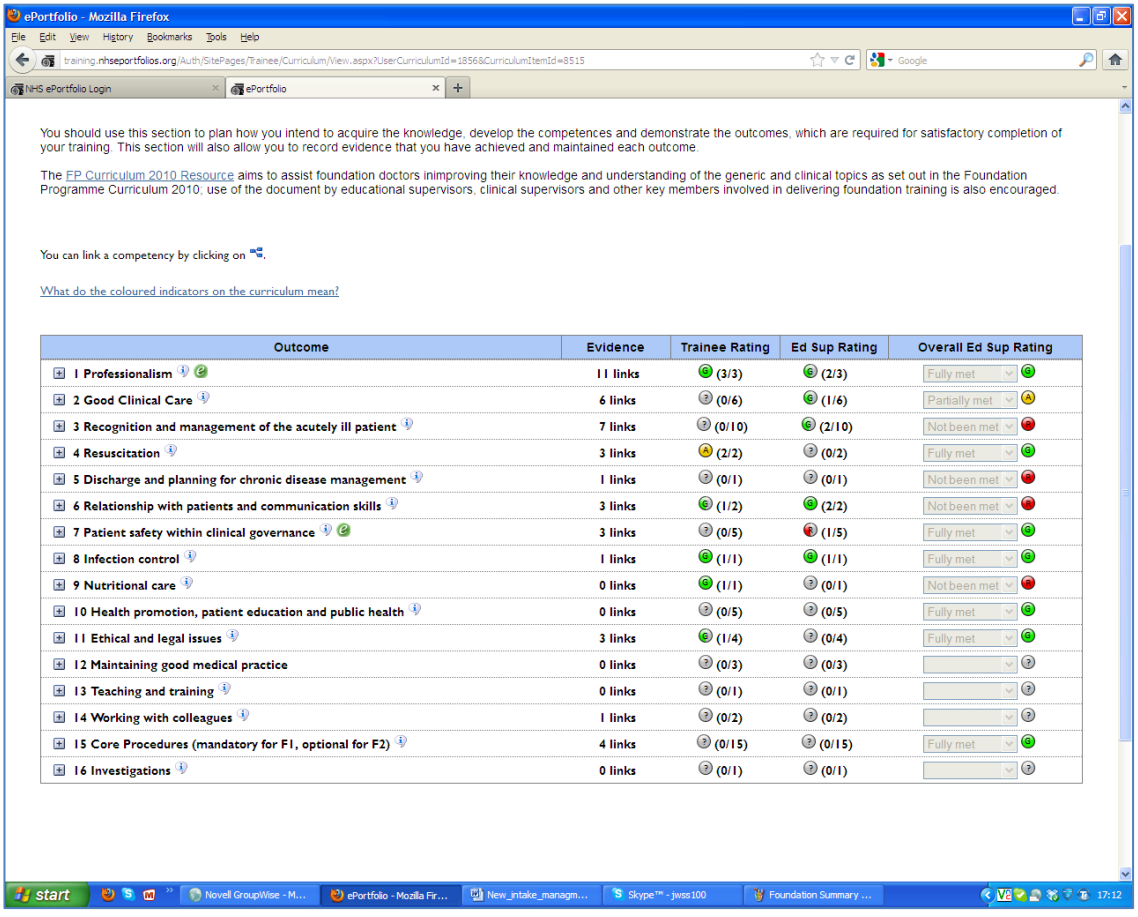

Figure 18. Screenshot of Foundation 2012 Curriculum and Associated Tools

The start of the 2012-13 training year in August saw further significant changes to curriculum, process and content. It is acknowledged that the workplace will be the

primary place of learning for both clinical and non-clinical skills, with the concepts of patient safety and personal development being central to all training. Whilst existing assessment tools (mini-CEX, DOPS, CBD) will continue to be used summatively, 2012 saw the introduction Supervised Learning Events (SLEs). These were intended to be frequent and unplanned events to be used in conjunction with assessments. Two of the crucial points SLEs are to address are the immediacy of feedback and encouragement of further structured development – both items which were seen to be truly deficient in the chapter above.

Self-assessment will continue to feature in the form of self-TAB (TAB replaced the Scottish MSF tool as well as the mini-PAT used elsewhere in the UK).Unfortunately despite the enhancements to the platform, self-TAB will not be specified for use in conjunction with other assessments as the literature suggests it would need to be, in order to be used most effectively. Similarly, no specific guidance is issued with regards to conducting a successful self-assessment (TAB).

7.4 Future Research

The self-assessment systematic review group formed in the months before the Foundation Programme and ePortfolio, began. This offered the opportunity to test the self-assessment review's questions against a large dataset ; however, the early years for both Foundation and ePortfolio saw sporadic use and consequently the data were not always informative with respect to the wider literature. Foundation and ePortfolio have evolved annually, and the Programme now has near-universal compliance and the ePortfolio sees extremely heavy traffic, making it an ideal data set for future research

The current NES ePortfolio contains a vast amount of educational data – over eight million rows – the vast majority of which is never used for research purposes. The potential of this data for research into assessment and the wider education of trainee medics, dentists, pharmacists and nurses is clearly enormous. Much has changed in the format and structure of the Foundation ePortfolio, but the content and processes have changed less. As the training year examined was much closer to the introduction of the

new programme there would be much benefit to re-examining the questions around self-assessment with data collected in a much more stable, established and accepted environment.

Future research should exploit the longitudinal potential of self-assessment data on electronic portfolios. The NES ePortfolio has data tracking some trainees from 2005 until present day. Currently there are no significant studies looking at longitudinal data in this area – and yet it is readily available. Longitudinal work could focus on comparing self-assessment scores with the scores and activities in other parts of the ePortfolio: how they compare with workplace based assessments, what impact they have with use of the educational log, how do they influence one's personal development plan, does the supervisors' report pick up on the results or consequent actions and (ultimately) is there evidence that can link the use of self-assessment to the certificate of completion. And rather than one-off isolated studies, these comparisons would reveal the results and outcomes of individuals and groups over time.

Variations within and between professional groups in how they use and view e-portfolios is also a topic that lies largely unexamined. Whilst there is anecdotal evidence that e-portfolios save time and money over physical copies this too has yet to be objectively measured. There are NES ePortfolio versions in use in dentistry, nursing, midwifery and pharmacy, therefore in theory, exploration and comparison of usage among these groups could be supported.

Finally, as items in the ePortfolio can be (and sometimes must be) linked, and learning is increasingly integrated within, the cognitive paths and behaviour of learners could readily be tracked so education could drive the development of the new media, rather than react to it. Linkage is also likely to be furthered with changing guidance about the use of routinely collected NHS staff data for research and evaluation, as there is an increasing emphasis on the use of administrative data to support public benefit.

In 2013 there has been the acceptance that ePortfolio is not best placed within a Special Health Board and invitations to the private and academic sector to form a joint venture are being sought. A partnership that alleviated the constraints of residing within the public sector, combined with academic expertise, could open a successful

but constricted product to provide a wealth of relational longitudinal educational research data to better analyse self-assessment, as well as a host of other educational subjects.

Dissemination

The author intends to disseminate the results of this thesis in a number of ways:

- Bring the findings to the relevant governance and educational content groups that oversee and influence the ePortfolio (e.g. COPMeD, UKFPO, Royal Colleges).
- Work with colleagues in the UK and internationally to raise awareness of the issues and explore future research for publication.
- Work with ePortfolio stakeholders to help them appreciate the enormous extent and potential of the evidence base contained in the application's databases, which is (largely) untapped.
- Present a paper to the Association of Medical Education in Europe 2014 evaluating the most recent training year's self-assessment data in terms of (a) whether the quartiles Kruger and Dunning observed are more readily discerned and (b) whether self-assessment can be seen to initiate other educational activity.

APPENDIX

Multi-Source Feedback

Assessor Name:

Assessor's Email:

Assessor Designation:

Assessor Location:

Trainee Name:
Trainee GMC:

Length of time working with trainee:

Clinical Care

The doctor is routinely able to take a structured history from the patients (carers)

Examples of unsatisfactory criteria:
- *Incomplete, inaccurate and confusing history taking from and communication with patients (carers)*
- *Fails to take into account the patients (carers) concerns, expectation or understanding*
- *May repeatedly upset patients (carers)*

Rating:

Highly Unsatisfactory	2	Unsatisfactory	4	Satisfactory	6	Highly satisfactory	***Cannot Evaluate***

Comments:

The doctor is able to promptly assess the acutely ill or collapsed patient

Examples of unsatisfactory criteria:
- *Unable to make an adequate medical assessment of airway, breathing, circulation*
- *Panics*
- *Does not call for help or advice appropriately*

Rating:

Highly Unsatisfactory	2	Unsatisfactory	4	Satisfactory	6	Highly satisfactory	***Cannot Evaluate***

Comments:

The doctor is able to appropriately manage and monitor the acutely ill or collapsed patient

Examples of unsatisfactory criteria:
- *Inappropriate administration of intravenous fluids/oxygen*
- *Does not initiate regular checking of unstable patients*
- *Unable to appreciate the urgency of the situation*
- *Does not initiate appropriate investigations*

Rating:

Highly Unsatisfactory	2	Unsatisfactory	4	Satisfactory	6	Highly satisfactory	***Cannot Evaluate***

Comments:

The doctor is able to prescribe safely and appropriate

Examples of unsatisfactory criteria:
- *Unaware of drug interactions*
- *Repeatedly prescribes inappropriately*
- *Does not adhere to protocols or guidelines*
- *Does not document prescribing clearing*
- *Does not discuss treatments and side effects with patients (carers)*
- *Unaware of safety issues in children, elderly, pregnancy*

Rating:

Highly Unsatisfactory	2	Unsatisfactory	4	Satisfactory	6	Highly satisfactory	***Cannot Evaluate***

Comments:

Global Rating

The doctor's overall performance is

Rating:

Highly Unsatisfactory	2	Unsatisfactory	4	Satisfactory	6	Highly satisfactory	***Cannot Evaluate***

Comments:

Any other general comments:

BIBLIOGRAPHY

Amery, J., Lapwood, S., 2004. A study into the educational needs of children's hospice doctors: a descriptive quantitative and qualitative survey. Palliative Medicine 18, 727 – 733.

Antonelli, M., 1997. Accuracy of second-year medical students' self-assessment of clinical skills. Acad Med 72, S63–65.

Archer, J., 2010. State of the science in health professional education: effective feedback. Medical Education 44, 101–108.

Austin, Z., Gregory, P., Galli, M., 2008. "I just don't know what I'm supposed to know": evaluating self-assessment skills of international pharmacy graduates in Canada. Res Social Adm Pharm 4, 115–124.

Austin, Z., Marini, A., DesRoches, B., 2005 Use of a learning portfolio for continuous professional development: A study of pharmacists in Ontario (Canada). Pharmacy Education 5, 175–181.

Avraamidou, L., 2003. Exploring the Influence of Web-Based Portfolio Development on Learning to Teach Elementary Science. JTATE 11, 415–442.

Avraamidou, L., Aiton, J., 2002. Student self-evaluation in clinical education. Radiol Technol 73, 415–422.

Baeten, M., Dochy, F., Struyven, K., 2008. Students' approaches to learning and assessment preferences in a portfolio-based learning environment. Instr Sci 36, 359–374.

Banister, S., Vannatta, R., Ross, C., 2006. Testing Electronic Portfolio Systems in a Teacher Education: Finding the Right Fit. Action in Teacher Education 27, 81–90.

Barnsley, L., Lyon, P., Ralston, S., Hibbert, E., Cunningham, I., Gordon, F., Field, M., 2004. Clinical skills in junior medical officers: a comparison of self-reported confidence and observed competence. Med Educ 38, 358–367.

Barrett, H., 2007. Researching electronic portfolios and learner engagement: The REFLECT initiative. Journal of Adolescent & Adult Literacy 50, 436–449.

Bartlett, A., Sherry, A., 2006. Two Views of Electronic Portfolios in Teacher Education: Non-Technology Undergraduates and Technology Graduate Students. International Journal of Instructional Media 33, 245–253.

Berwick, D., 2005. Broadening the view of evidence-based medicine. Qual Saf Health Care 14, 315–316.

Biernat, K., Simpson, D., Duthie, E., Bragg, D., London, R., 2003. Primary care residents self assessment skills in dementia. Adv Health Sci Educ Theory Pract 8, 105–110.

Biran, L., 1991. Self-assessment and learning through GOSCE (group objective structured clinical examination). Med Educ 25, 475–479.

Bowers, S., Jinks, A., 2004. Issues surrounding professional portfolio development for nurses. Br J Nurs 13, 155–159.

Bradley, E., Dolovich, L., Austin, Z., 2007. Comparison of self, physician, and simulated patient ratings of pharmacist performance in a family practice simulator. J Interprof Care 21, 129–140.

Brewster, L., Risucci, D., Joehl, R., Littooy, F., Temeck, B., Blair, P., Sachdeva, A., 2008. Comparison of resident self-assessments with trained faculty and standardized patient assessments of clinical and technical skills in a structured educational module. Am. J. Surg. 195, 1–4.

Brinkman, W., Geraghty, S., Lanphear, B., Khoury, J., Gonzalez del Rey, J., Dewitt, T., Britto, M., 2007. Effect of multisource feedback on resident communication skills and professionalism: a randomized controlled trial. Arch Pediatr Adolesc Med 161, 44–49.

Bryan, R., Krych, A., Carmichael, S., Viggiano, T., Pawlina, W., 2005. Assessing professionalism in early medical education: experience with peer evaluation and self-evaluation in the gross anatomy course. Ann. Acad. Med. Singap. 34, 486–491.

Buckley, S., Coleman, J., Davison, I., Khan, K., Zamora, J., Malick, S., Morley, D., Pollard, D., Ashcroft, T., Popovic, C., Sayers, J., 2009. The educational effects of portfolios on undergraduate student learning: a Best Evidence Medical Education (BEME) systematic review. BEME Guide No. 11. Med Teach 31, 282–298.

Bullock, A., Hassell, A., Markham, W., Wall, D., Whitehouse, A., 2009. How ratings vary by staff group in multi-source feedback assessment of junior doctors. Medical Education 43, 516–520.

Butterfield, B., Metcalfe, J., 2001. Errors committed with high confidence are hypercorrected. J Exp Psychol Learn Mem Cogn 27, 1491–1494.

Campbell, C., Parboosingh, J., Gondocz, S., Babitskaya, G., Lindsay, E., De Guzman, R., Klein, L., 1996. Study of physicians' use of a software program to create a portfolio of their self-directed learning. Acad Med 71, S49–51.

Campbell, C., Parboosingh, J., Gondocz, T., Babitskaya, G., Pham, B., 1999. Study of the factors influencing the stimulus to learning recorded by physicians keeping a learning portfolio. Journal of Continuing Education in the Health Professions 19, 16–24.

Carney, J., Jay, J., 2002. Translating Theory into Practice: The Dilemmas of Teacher Portfolios.

http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED462438 [accessed 17/09/11]

Carraccio, C., Englander, R., 2004. Evaluating competence using a portfolio: a literature review and web-based application to the ACGME competencies. Teaching and Learning in Medicine 16, 381–387.

CASP UK, Critical Appraisal Skills Programme. http://www.casp-uk.net/ [accessed 4/2/14]

Caverzagie, K., Shea, J., Kogan, J., 2008. Resident identification of learning objectives after performing self-assessment based upon the ACGME core competencies. J Gen Intern Med 23, 1024–1027.

Chabeli, M., 2002. Portfolio assessment and evaluation: implications and guidelines for clinical nursing education. Curationis 25, 4–9.

Challis, M., 1999. AMEE Medical Education Guide No.11 (revised): Portfolio-based learning and assessment in medical education. Medical Teacher 21, 370–386.

Challis, M., Mathers, N., Howe, A., Field, N., 1997. Portfolio-based learning: continuing medical education for general practitioners—a mid-point evaluation. Medical Education 31, 22–26.

Chang, C., 2001. A study on the evaluation and effectiveness analysis of web-based learning portfolio (WBLP). British Journal of Educational Technology 32, 435–458.

Chur-Hansen, A., 2000. Medical students' essay-writing skills: criteria-based self- and tutor-evaluation and the role of language background. Med Educ 34, 194–198.

Cise, J., Wilson, C., Thie, M., 2004. A qualitative tool for critical thinking skill development. Nurse Educ 29, 147–151.

Clegg, S., 2005. Evidence-based practice in educational research: a critical realist critique of systematic review. British Journal of Sociology of Education 26, 415–428.

Clegg, S., Hudson, A., Mitchell, A., 2005. The Personal Created through Dialogue: Enhancing Possibilities through the Use of New Media. ALT-J: Research in Learning Technology 13, 3–15.

Coffey, A., 2005. The clinical learning portfolio: a practice development experience in gerontological nursing. J Clin Nurs 14, 75–83.

Coleman, H., Morris, D., Norton, R., 2006. Developing Multicultural Counseling Competence through the Use of Portfolios. Journal of Multicultural Counseling and Development 34, 27.

Collins, J., 2010. Foundation for Excellence. An evaluation of the Foundation

Programme. Medical Education England.

Colliver, J., Verhulst, S., Barrows, H., 2005. Self-assessment in medical practice: a further concern about the conventional research paradigm. Teach Learn Med 17, 200–201.

Colthart, I., Bagnall, G., Evans, A., Allbutt, H., Haig, A., Illing, J., McKinstry, B., 2008. The effectiveness of self-assessment on the identification of learner needs, learner activity, and impact on clinical practice: BEME Guide no. 10. Medical Teacher 30, 124–145.

Cope, M., Baker, H., Foster, R., Boisvert, C., 2007. Relationships Between Clinical Rotation Subscores, COMLEX-USA Examination Results, and School-Based Performance Measures. J Am Osteopath Assoc 107, 502–510.

Cotterill, S., Aiton, J., Bradley, P., Hammond, G., McDonald, A., Struthers, J., Whiten, S., 2006. A Flexible Component-Based ePortfolio: Embedding in the Curriculum. Handbook of research on ePortfolios 292–304.

Craig, E., 2007. Changing paradigms: managed learning environments and Web 2.0. Campus-Wide Information Systems 24, 152–161.

Crawford, M., Kiger, A., 1998. Development through self-assessment: strategies used during clinical nursing placements. J Adv Nurs 27, 157–164.

Cross, M., White, P., 2004. Personal development plans: the Wessex experience. Education for Primary Care 15, 205–212.

Dagley, V., Berrington, B., 2005. Learning from an evaluation of an electronic portfolio to support general practitioners' personal development planning, appraisal and revalidation. Education for Primary Care 16, 567–574.

Dauphinee, W., Wood-Dauphinee, S., 2004. The need for evidence in medical education: the development of best evidence medical education as an opportunity to inform, guide, and sustain medical education research. Acad Med 79, 925–930.

Davies, P., 1999. What is Evidence-based Education? British Journal of Educational Studies 47, 108–121.

Davies, P., 2000. The relevance of systematic reviews to educational policy and practice. Oxford Review of Education 365–378.

Davis, D., Mazmanian, P., Fordis, M., Van Harrison, R., Thorpe, K., Perrier, L., 2006. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. JAMA 296, 1094–1102.

Davis, J., 2002. Comparison of faculty, peer, self, and nurse assessment of obstetrics and gynecology residents. Obstet Gynecol 99, 647–651.

Davis, M., Ponnamperuma, G., 2010. Examiner perceptions of a portfolio assessment process. Med Teach 32, e211–215.

Deketelaere, A., Degryse, J., De Munter, A., De Leyn, P., 2009. Twelve tips for successful e-tutoring using electronic portfolios. Med Teach 31, 497–501.

Dekker, H., Driessen, E., Ter Braak, E., Scheele, F., Slaets, J., Van Der Molen, T., Cohen-Schotanus, J., 2009. Mentoring portfolio use in undergraduate and postgraduate medical education. Med Teach 31, 903–909.

Department of Trade and Industry, 2003. The Working Time (Amendment) Regulations 2003.

Dixon-Woods, M., Agarwal, S., Jones, D., Young, B., Sutton, A., 2005. Synthesising qualitative and quantitative evidence: a review of possible methods. Journal of health services research & policy 10, 45.

Dixon-Woods, M., Bonas, S., Booth, A., Jones, D., Miller, T., Sutton, A., Shaw, R., Smith, J., Young, B., 2006. How can systematic reviews incorporate qualitative research? A critical perspective. Qualitative Research 6, 27–44.

Dorn, C., Sabol, F., 2006. The Effectiveness and Use of Digital Portfolios for the Assessment of Art Performances in Selected Secondary Schools. Studies in Art Education: A Journal of Issues and Research in Art Education 47, 344–362.

Dornan, T., 2008. Self-assessment in CPD: lessons from the UK undergraduate and postgraduate education domains. J Contin Educ Health Prof 28, 32–37.

Dornan, T., Carroll, C., Parboosingh, J., 2002. An electronic learning portfolio for reflective continuing professional development. Med Educ 36, 767–769.

Dornan, T., Lee, C., Stopford, A., Hosie, L., Maredia, N., Rector, A., 2005. Rapid application design of an electronic clinical skills portfolio for undergraduate medical students. Comput Methods Programs Biomed 78, 25–33.

Dornan, T., Maredia, N., Hosie, L., Lee, C., Stopford, A., 2003. A web-based presentation of an undergraduate clinical skills curriculum. Med Educ 37, 500–508.

Driessen, E., Muijtjens, A., van Tartwijk, J., van der Vleuten, C., 2007. Web- or paper-based portfolios: is there a difference? Med Educ 41, 1067–1073.

Driessen, E., van Tartwijk, J., Dornan, T., 2008. The self critical doctor: helping students become more reflective. BMJ 336, 827–830.

Driessen, E., van Tartwijk, J., van der Vleuten, C., Wass, V., 2007. Portfolios in medical education: why do they meet with mixed success? A systematic review. Med Educ 41, 1224–1233.

Duffy, F., Holmboe, E., 2006. Self-assessment in Lifelong Learning and Improving Performance in Practice. JAMA: The Journal of the American Medical Association 296, 1137 –1139.

Dunning, D., 2006. Strangers to ourselves. The Psychologist 19, 600–603.

Duque, G., Finkelstein, A., Roberts, A., Tabatabai, D., Gold, S., Winer, L., 2006. Learning while evaluating: the use of an electronic evaluation portfolio in a geriatric medicine clerkship. BMC Med Educ 6, 4.

Edwards, R., Kellner, K., Sistrom, C., Magyari, E., 2003. Medical student self-assessment of performance on an obstetrics and gynecology clerkship. Am. J. Obstet. Gynecol. 188, 1078–1082.

Ehrlinger, J., Dunning, D., 2003. How chronic self-views influence (and potentially mislead) estimates of performance. J Pers Soc Psychol 84, 5–17.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., Kruger, J., 2008. Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. Organizational Behavior and Human Decision Processes 105, 98–121.

Elizur, A., Kretsch, R., Spaizer, N., Sorek, Y., 1994. Self-evaluation of psychotherapeutic competence. Br J Med Psychol 67 (Pt 3), 231–235.

Elliott, N., Higgins, A., 2005. Self and peer assessment ? does it make a difference to student group work? Nurse Education in Practice 5, 40–48.

Ellis, J., Teasdale, D., Cotterill, S., Thomason, J., 2010. Is a generic UK e-portfolio for dentistry desirable and achievable? European Journal of Dental Education 14, 254–256.

Epstein, R., 2007. Assessment in Medical Education. New England Journal of Medicine 356, 387–396.

Epstein, R., Siegel, D., Silberman, J., 2008. Self-monitoring in clinical practice: a challenge for medical educators. J Contin Educ Health Prof 28, 5–13.

Ericson, D., Christersson, C., Manogue, M., Rohlin, M., 1997. Clinical guidelines and self-assessment in dental education. Eur J Dent Educ 1, 123–128.

Eva, K., Cunnington, J., Reiter, H., Keane, D., Norman, G., 2004. How can I know what I don’t know? Poor self assessment in a well-defined domain. Adv Health Sci Educ Theory Pract 9, 211–224.

Eva, K., Regehr, G., 2005. Self-assessment in the health professions: a reformulation and research agenda. Acad Med 80, S46–54.

Eva, K., Regehr, G., 2008. “I’ll never play professional football” and other fallacies of

self-assessment. J Contin Educ Health Prof 28, 14–19.

Eva, W., Regehr, G., 2007. Knowing when to look it up: a new conception of self-assessment ability. Acad Med 82, S81–84.

Evans, A., Aghabeigi, B., Leeson, R., O'Sullivan, C., Eliahoo, J., 2002. Are we really as good as we think we are? Ann R Coll Surg Engl 84, 54–56.

Evans, A., Leeson, R., Newton John, T., Petrie, A., 2005. The influence of self-deception and impression management upon self-assessment in oral surgery. British Dental Journal 198, 765–769.

Evans, A., Leeson, R., Petrie, A., 2007. Reliability of peer and self-assessment scores compared with trainers' scores following third molar surgery. Med Educ 41, 866–872.

Evans, D., 2003. Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. Journal of Clinical Nursing 12, 77–84.

Farnill, D., Hayes, S., Todisco, J., 1997. Interviewing skills: self-evaluation by medical students. Med Educ 31, 122–127.

Fincher, R., Lewis, L., 1994. Learning, experience, and self-assessment of competence of third-year medical students in performing bedside procedures. Acad Med 69, 291–295.

Fincher, R., Lewis, L., Kuske, T., 1993. Relationships of interns' performances to their self-assessments of their preparedness for internship and to their academic performances in medical school. Acad Med 68, S47–50.

Fitzgerald, J., Gruppen, L., White, C., 2000. The influence of task formats on the accuracy of medical students' self-assessments. Acad Med 75, 737–741.

Fitzgerald, J., White, C., Gruppen, L., 2003. A longitudinal study of self-assessment accuracy. Med Educ 37, 645–649.

Flores-Mateo, G., Argimon, J., 2007. Evidence based practice in postgraduate healthcare education: A systematic review. BMC Health Services Research 7, 119.

Fox, R., Ingham Clark, C., Scotland, A., Dacre, J., 2000. A study of pre-registration house officers' clinical skills. Med Educ 34, 1007–1012.

Francke, A., Garssen, B., Abu-Saad, H., 1995. Determinants of changes in nurses' behaviour after continuing education: a literature review. Journal of Advanced Nursing 21, 371–377.

Frye, A., Richards, B., Bradley, E., Philp, J., 1992. The consistency of students' self-assessments in short-essay subject matter examinations. Med Educ 26, 310–316.

Fung, M., Walker, M., Fung, K., Temple, L., Lajoie, F., Bellemare, G., Bryson, S., 2000. An Internet-based learning portfolio in resident education: the KOALA™ multicentre programme. Medical Education 34, 474–479.

Galbraith, R., Hawkins, R., Holmboe, E., 2008. Making self-assessment more effective. J Contin Educ Health Prof 28, 20–24.

Garrett, B., Jackson, C., 2006. A mobile clinical e-portfolio for nursing and medical students, using wireless personal digital assistants (PDAs). Nurse Educ Today 26, 647–654.

Garrett, B.M., Jackson, C., 2006. A mobile clinical e-portfolio for nursing and medical students, using wireless personal digital assistants (PDAs). Nurse Education in Practice 6, 339–346.

General Medical Council, 2006. Good Medical Practice. Standards guidance for doctors. http://www.gmc-uk.org/publications/standards_guidance_for_doctors.asp#gmp [accessed 19/10/11]

General Medical Council, 2010. Workplace Based Assessment: A guide for implementation. http://www.gmc-uk.org/Workplace_based_assessment_31381027.pdf [accessed 19/10/11]

General Medical Council, 2011a. Revalidation: the way ahead. 20 questions the GMC wants you to answer. http://www.gmc-uk.org/news/5881.asp [accessed 24/09/11]

General Medical Council, 2011b. Tomorrow's Doctors online (2009).

General Medical Council, Guidance on CPD. http://www.gmc-uk.org/education/continuing_professional_development/cpd_guidance.asp [accessed 20/4/13]

General Medical Council, Standards for curricula and assessment systems. http://www.gmc-uk.org/education/postgraduate/standards_for_curricula_and_assessment_systems.asp [accessed 19/10/11]

General Medical Council, Supporting information for appraisal and revalidation. http://www.gmc-uk.org/doctors/revalidation/revalidation_information.asp [accessed 24/11/12]

Gordon, M., 1991. A review of the validity and accuracy of self-assessments in health professions training. Acad Med 66, 762–769.

Gordon, M., 1992. Self-assessment programs and their implications for health professions training. Academic Medicine 67.

Gordon, M., 1997. Cutting the Gordian knot: a two-part approach to the evaluation

and professional development of residents. Acad Med 72, 876–880.

Gray, L., 2008. “Effective Practice with e-Portfolios” publication.

Greenhalgh, T., Toon, P., Russell, J., Wong, G., Plumb, L., Macfarlane, F., 2003. Transferability of principles of evidence based medicine to improve educational quality: systematic review and case study of an online course in primary health care. BMJ 326, 142 –145.

Gruppen, L., White, C., Fitzgerald, J., Grum, C., Woolliscroft, J., 2000. Medical students’ self-assessments and their allocations of learning time. Acad Med 75, 374–379.

Guskey, T., 2007. Multiple Sources of Evidence: An Analysis of Stakeholders’ Perceptions of Various Indicators of Student Learning. Educational Measurement: Issues and Practice 26, 19–27.

Haffling, A., Beckman, A., Pahlmblad, A., Edgren, G., 2010. Students’ reflections in a portfolio pilot: highlighting professional issues. Med Teach 32, e532–540.

Hammersley, M., 2005. Is the evidence-based practice movement doing more good than harm? Reflections on Iain Chalmers’ case for research-based policy making and practice. Evidence &# 38; Policy: A Journal of Research, Debate and Practice 1, 85–100.

Hammick, M., Haig, A., 2007. The Best Evidence Medical Education Collaboration: processes, products and principles. The Clinical Teacher 4, 42–45.

Hammond, D., Buckendahl, C., 2006. Do portfolio assessments have a place in dental licensure? Against the proposition. J Am Dent Assoc 137, 30, 32, 34 passim.

Harden, R., Grant, J., Buckley, G., Hart, I., 2000. Best Evidence Medical Education. Adv Health Sci Educ Theory Pract 5, 71–90.

Harrington, J., Murnaghan, J., Regehr, G., 1997. Applying a relative ranking model to the self-assessment of extended performances. Adv Health Sci Educ Theory Pract 2, 17–25.

Hartman, S., Nelson, M., 1992. What we say and what we do: self-reported teaching behavior versus performances in written simulations among medical school faculty. Acad Med 67, 522–527.

Hartzler, B., Baer, J., Dunn, C., Rosengren, D., Wells, E., 2007. What is Seen Through the Looking Glass: The Impact of Training on Practitioner Self-Rating of Motivational Interviewing Skills. Behavioural and Cognitive Psychotherapy 35, 431.

Hauge, T., 2006. Portfolios and ICT as means of professional learning in teacher education. Studies in Educational Evaluation 32, 23–36.

Henderson, P., Johnson, M., 2002. An innovative approach to developing the reflective

skills of medical students. BMC Med Educ 2, 4.

Hennessy, S., Howes, A., 2004. Using ePortfolios to assess the Reflective Capabilities of Medical Students 1–10.

Herbert, W., McGaghie, W., Droegemueller, W., Riddle, M., Maxwell, K., 1990. Student evaluation in obstetrics and gynecology: self- versus departmental assessment. Obstet Gynecol 76, 458–461.

Hodges, B., Regehr, G., Martin, D., 2001. Difficulties in recognizing one's own incompetence: novice physicians who are unskilled and unaware of it. Acad Med 76, S87–89.

Hoppe, R., Farquhar, L., Henry, R., Stoffelmayr, B., 1990. Residents' attitudes towards and skills in counseling: using undetected standardized patients. J Gen Intern Med 5, 415–420.

Horner, A., Cotterill, S., Ingraham, B., Thompson, J., Gill, S., Ayestaran, H., Webster, D., Ollerenshaw, B., Mcdonald, A., Taylor, L., Wilson, R., Quentin-Baxter, M., Hopkins, P., n.d. EPICS:Outcomes of a regional ePortfolio initiative to support lifelong learning — EIfEL.

Hough, A.,"Disastrous" £11.4bn NHS IT programme to be abandoned. The Telegraph. http://www.telegraph.co.uk/health/healthnews/8780566/Disastrous-11.4bn-NHS-IT-programme-to-be-abandoned.html [accessed 19/10/11]

Hrisos, S., Illing, J., Burford, B., 2008. Portfolio learning for foundation doctors: early feedback on its use in the clinical workplace. Med Educ 42, 214–223.

Humphry, R., Geissinger, S., 1992. Self-Rating as an Evaluation Tool Following Continuing Professional Education. Occupational Therapy Journal of Research 12, 111–22.

Hussain, W., Hafiji, J., Stanley, A., Khan, K., 2008. Dermatology and junior doctors: an evaluation of education, perceptions and self-assessed competencies. Br. J. Dermatol. 159, 505–506.

Hutchinson, L., 1999. Evaluating and researching the effectiveness of educational interventions. BMJ 318, 1267–1269.

Indulski, J., Boczkowski, A., 1999. Self-assessment of competence in public health management as a measure of effectiveness of postgraduate training. Int J Occup Med Environ Health 12, 15–27.

Jacob, J., Ostchega, Y., Grady, C., Gallaway, L., Kish, J., 1990. Self-assessed learning needs of oncology nurses caring for individuals with HIV-related disorders. A national survey. Cancer Nurs 13, 246–255.

James, D., 1992. A unique manual for self-assessment by dental practitioners. Quintessence Int 23, 701–704.

Jarvis, R., O'Sullivan, P., McClain, T., Clardy, J., 2004. Can One Portfolio Measure the Six ACGME General Competencies? Acad Psychiatry 28, 190–196.

Jasper, M., Fulton, J., 2005. Marking criteria for assessing practice-based portfolios at masters' level. Nurse Education Today 25, 377–389.

Jensen, G., Saylor, C., 1994. Portfolios and Professional Development in the Health Professions. Evaluation & the Health Professions 17, 344 –357.

Johnson, D., Cujec, B., 1998. Comparison of self, nurse, and physician assessment of residents rotating through an intensive care unit. Crit. Care Med. 26, 1811–1816.

Kaiser, S., Bauer, J., 1995. Checklist self-evaluation in a standardized patient exercise. Am. J. Surg. 169, 418–420.

Keim, K., Gates, G., Johnson, C., 2001. Dietetics professionals have a positive perception of professional development. J Am Diet Assoc 101, 820–824.

Kirkpatrick, D., Evaluation of training. McGraw Hill, New York, pp. 87–112.

Kjaer, N., Maagaard, R., Wied, S., 2006. Using an online portfolio in postgraduate training. Med Teach 28, 708–712.

Kramer, A., Zuithoff, P., Jansen, J., Tan, L., Grol, R., Vleuten, C., 2006. Growth of Self-Perceived Clinical Competence in Postgraduate Training for General Practice and its Relation to Potentially Influencing Factors. Advances in Health Sciences Education 12, 135–145.

Kruger, J., Dunning, D., 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of personality and social psychology 77, 1121.

Kuiper, RA., Pesut, D., 2004. Promoting cognitive and metacognitive reflective reasoning skills in nursing practice: self-regulated learning theory. J Adv Nurs 45, 381–391.

Lane, J., Gottlieb, R., 2004. Improving the interviewing and self-assessment skills of medical students: is it time to readopt videotaping as an educational tool? Ambul Pediatr 4, 244–248.

Lang, T., 2004. The value of systematic reviews as research activities in medical education. Acad Med 79, 1067–1072.

Langendyk, V., 2006. Not knowing that they do not know: self-assessment accuracy of third-year medical students. Medical Education 40, 173–179.

Larsen, T., Jeppe-Jensen, D., 2008. The introduction and perception of an OSCE with an element of self- and peer-assessment. Eur J Dent Educ 12, 2–7.

Leach, D., 2002. Competence is a habit. JAMA 287, 243–244.

Learman, L., Autry, A., O'Sullivan, P., 2008. Reliability and validity of reflection exercises for obstetrics and gynecology residents. Am. J. Obstet. Gynecol. 198, 461.e1–8; discussion 461.e8–10.

Ledoux, M., Mchenry, N., 2006. Electronic Portfolio Adoption for Teacher Education Candidates. Early Childhood Education Journal 34, 103–116.

Leisnert, L., Mattheos, N., 2006. The interactive examination in a comprehensive oral care clinic: a three-year follow up of students' self-assessment ability. Med Teach 28, 544–548.

Leopold, S., Morgan, H., Kadel, N., Gardner, G., Schaad, D., Wolf, F., 2005. Impact of educational intervention on confidence and competence in the performance of a simple surgical task. J Bone Joint Surg Am 87, 1031–1037.

Levinson, W., Gordon, G., Skeff, K., 1990. Retrospective Versus Actual Pre-Course Self-Assessments. Evaluation & the Health Professions 13, 445–452.

Lewis, K., Baker, R., 2007. The development of an electronic educational portfolio: an outline for medical education professionals. Teach Learn Med 19, 139–147.

Lin, K., Yang, S., Hung, J., Wang, D., 2006. Web-Based Appreciation and Peer-Assessment for Visual-Art Education. International Journal of Distance Education Technologies 4, 5–14.

Lockyer, J., 2003. Multisource feedback in the assessment of physician competencies. Journal of Continuing Education in the Health Professions 23, 4–12.

Lurie, S., Mooney, C., Lyness, J., 2009. Measurement of the General Competencies of the Accreditation Council for Graduate Medical Education: A Systematic Review. Academic Medicine 84, 301–309.

Lynch, D., Swing, S., Horowitz, S., Holt, K., Messer, J., 2004. Assessing practice-based learning and improvement. Teach Learn Med 16, 85–92.

Maidment, Y., Rennie, J., Thomas, M., 2006. Revalidation of general dental practitioners in Scotland: The results of a pilot study Part 1 - feasibility of operation. Br Dent J 200, 399–402.

Mandel, L., Goff, B., Lentz, G., 2005. Self-assessment of resident surgical skills: is it feasible? Am. J. Obstet. Gynecol. 193, 1817–1822.

Mann, K., 2011. Theoretical perspectives in medical education: past experience and

future possibilities. Medical Education 45, 60–68.

Martin, D., Regehr, G., Hodges, B., McNaughton, N., 1998. Using videotaped benchmarks to improve the self-assessment ability of family practice residents. Acad Med 73, 1201–1206.

Mason, R., 2006. Learning technologies for adult continuing education. Studies in Continuing Education 28, 121–133.

Mathers, N., Challis, M., Howe, A., Field, N., 1999. Portfolios in continuing medical education--effective and efficient? Med Educ 33, 521–530.

Mattheos, N., Nattestad, A., Falk-Nilsson, E., Attström, R., 2004. The interactive examination: assessing students' self-assessment ability. Med Educ 38, 378–389.

Maudsley, G., Strivens, J., 2000. Promoting professional knowledge, experiential learning and critical thinking for medical students. Medical Education 34, 535–544.

Mayer, D., 2004. Essential evidence-based medicine. Cambridge University Press.

Mays, N., Pope, C., Popay, J., 2005. Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. Journal of health services research & policy 10, 6.

McCord, E., Smorowski-Garcia, K., Doughty, A., 1997. Assessment at one school of students' abilities and confidence in diabetic patients' education. Acad Med 72, 1116–1118.

McCready, T., 2007. Portfolios and the assessment of competence in nursing: A literature review. International journal of nursing studies 44, 143–151.

McMullan, M., Endacott, R., Gray, M., Jasper, M., Miller, C., Scholes, J., Webb, C., 2003. Portfolios and assessment of competence: a review of the literature. J Adv Nurs 41, 283–294.

Melville, C., Rees, M., Brookfield, D., Anderson, J., 2004. Portfolios for assessment of paediatric specialist registrars. Med Educ 38, 1117–1125.

Metcalfe, J., Finn, B., 2011. People's hypercorrection of high-confidence errors: did they know it all along? J Exp Psychol Learn Mem Cogn 37, 437–448.

Miller, A., Archer, J., 2010. Impact of workplace based assessment on doctors' education and performance: a systematic review. BMJ 341, c5064–c5064.

Miller, G., 1990. The assessment of clinical skills/competence/performance. Academic Medicine 65.

Miller, P., 1999. The agreement of peer assessment and self-assessment of learning

processes in problem-based learning. Journal of Physical Therapy Education HighBeam Research.

Miller, S., 2005. American board of medical specialties and repositioning for excellence in lifelong learning: Maintenance of certification. Journal of Continuing Education in the Health Professions 25, 151–156.

Millis, S., Jain, S., Eyles, M., Tulsky, D., Nadler, S., Bradley, P., Elovic, E., DeLisa, J., 2002. Assessing physicians' interpersonal skills: do patients and physicians see eye-to-eye? Am J Phys Med Rehabil 81, 946–951.

Minter, R., Gruppen, L., Napolitano, K., Gauger, P., 2005. Gender differences in the self-assessment of surgical residents. Am. J. Surg. 189, 647–650.

Moorthy, K., Munz, Y., Adams, S., Pandey, V., Darzi, A., 2006. Self-assessment of performance among surgical trainees during simulated procedures in a simulated operating theater. Am. J. Surg. 192, 114–118.

Moyer, J., 2002. The APNG(c): A preliminary look at credentialing nurses through portfolio review. Newborn and Infant Nursing Reviews 2, 254–258.

Mullan, P., Blitz, S., Stross, J., 1992. Faculty expectations and primary care residents' perceptions concerning residents' growth in competence at one medical school. Acad Med 67, 113–117.

Murdock, J., Neafsey, P., 1995. Self-efficacy measurements: an approach for predicting practice outcomes in continuing education? J Contin Educ Nurs 26, 158–165.

Murray, C., Currant, N., 2006. E-portfolios along the Lifelong Learning Cycle: Differences between Use, Pedagogy and Context., in: Sixth International Conference on Advanced Learning Technologies, 2006. Presented at the Sixth International Conference on Advanced Learning Technologies, 2006, pp. 491–493.

NHS Education for Scotland. http://www.nes.scot.nhs.uk/ [accessed 10/19/11].

NHS ePortfolio Login https://www.nhseportfolios.org/Anon/Login/Login.aspx?ReturnUrl=%2fAuth%2fCommon%2fPages%2fSelectRole.aspx [accessed 10/19/11].

National Institute for Health and Care Excellence (NICE), 2013. How to change practice: understand, identify and overcome barriers to change (Guidance/Implementation).

Norcini, J., Lipner, R., Downing, S., 1996. How meaningful are scores on a take-home recertification examination? Acad Med 71, S71–73.

Norcini, J., McKinley, D., 2007. Assessment methods in medical education. Teaching and Teacher Education 23, 239–250.

Norman, G., Shannon, S., Marrin, M., 2004. The need for needs assessment in continuing medical education. BMJ 328, 999–1001.

Nursing and Midwifery Council, 2010. Safeguarding health and wellbeing.

O'Brien, M., Brown, J., Ryland, I., Shaw, N., Chapman, T., Gillies, R., Graham, D., 2006. Exploring the views of second-year Foundation Programme doctors and their educational supervisors during a deanery-wide pilot Foundation Programme. Postgrad Med J 82, 813–816.

O'Sullivan, P., Reckase, M., McClain, T., Savidge, M., Clardy, J., 2004. Demonstration of portfolios to assess competency of residents. Adv Health Sci Educ Theory Pract 9, 309–323.

Oetting, T., Lee, A., Beaver, H., Johnson, A., Boldt, H., Olson, R., Carter, K., 2006. Teaching and assessing surgical competency in ophthalmology training programs. Ophthalmic Surg Lasers Imaging 37, 384–393.

Pandey, V., Wolfe, J., Black, S., Cairols, M., Liapis, C., Bergqvist, D., 2008. Self-assessment of technical skill in surgery: the need for expert feedback. Ann R Coll Surg Engl 90, 286–290.

Paradise, J., Finkel, M., Beiser, A., Berenson, A., Greenberg, D., Winter, M., 1997. Assessments of girl's genital findings and the likelihood of sexual abuse: agreement among physicians self-rated as skilled. Arch Pediatr Adolesc Med 151, 883–891.

Parker, R., Alford, C., Passmore, C., 2004. Can family medicine residents predict their performance on the in-training examination? Fam Med 36, 705–709.

Pearson, D., Heywood, P., 2004. Portfolio use in general practice vocational training: a survey of GP registrars. Med Educ 38, 87–95.

Pelayo, M., Cebrián, D., Areosa, A., Agra, Y., Izquierdo, J., Buendía, F., n.d. Effects of online palliative care training on knowledge, attitude and satisfaction of primary care physicians. BMC Fam Pract 12, 37–37.

Petersen, S., 1999. Time for evidence based medical education. BMJ 318, 1223–1224.

Pierre, R., Wierenga, A., Barton, M., Thame, K., Branday, J., Christie, C., 2005. Student self-assessment in a paediatric objective structured clinical examination. West Indian Med J 54, 144–148.

Pitts, J., Coles, C., Thomas, P., 2001. Enhancing reliability in portfolio assessment: "shaping" the portfolio. Med Teach 23, 351–356.

Pitts, J., Coles, C., Thomas, P., Smith, F., 2002. Enhancing reliability in portfolio assessment: discussions between assessors. Med Teach 24, 197–201.

Price, B., 2005. Self-assessment and reflection in nurse education. Nurs Stand 19, 33–37.

Prideaux, D., 2002. Researching the outcomes of educational interventions: a matter of design. BMJ 324, 126.

Raelin, J., 2001. Public Reflection as the Basis of Learning. Management Learning 32, 11–30.

Rassin, M., Silner, D., Ehrenfeld, M., 2006. Departmental portfolio in nursing - An advanced instrument. Nurse Educ Pract 6, 55–60.

Redish, T., Webb, L., Jiang, B., 2006. Design and Implementation of a Web-Based Portfolio for Aspiring Educational Leaders: A Comprehensive, Evidence-Based Model. Journal of Educational Technology Systems 34, 283–295.

Rees, C., Shepherd, M., 2005. Students' and assessors' attitudes towards students' self-assessment of their personal and professional behaviours. Med Educ 39, 30–39.

Reisine, S., 1996. An overview of self-reported outcome assessment in dental research. J Dent Educ 60, 488–493.

Reiter, H., Eva, K., Hatala, R., Norman, G., 2002. Self and peer assessment in tutorials: application of a relative-ranking model. Acad Med 77, 1134–1139.

Richardson, A., 1998. Personal professional profiles. Nurs Stand 12, 35–40.

Rosenberg, M., Watson, K., Paul, J., Miller, W., Harris, I., Valdivia, T., 2001. Development and implementation of a web-based evaluation system for an internal medicine residency program. Acad Med 76, 92–95.

Rosewell, J., 2012. A speculation on the possible use of badges for learning at the UK Open University (Conference Item).

Rudy, D., Fejfar, M., Griffith, C., Wilson, J., 2001. Self- and Peer Assessment in a First-Year Communication and Interviewing Course. Evaluation & the Health Professions 24, 436 –445.

Russell, J., Elton, L., Swinglehurst, D., Greenhalgh, T., 2006. Using the online environment in assessment for learning: a case-study of a web-based course in primary care. Assessment & Evaluation in Higher Education 31, 465–478.

Ryland, I., Brown, J., O'Brien, M., Graham, D., Gillies, R., Chapman, T., Shaw, N., 2006. The portfolio: how was it for you? Views of F2 doctors from the Mersey Deanery Foundation Pilot. Clin Med 6, 378–380.

Sargeant, J., Mann, K., van der Vleuten, C., Metsemakers, J., 2008. "Directed" self-assessment: practice and feedback within a social context. J Contin Educ Health Prof

28, 47–54.

Schmidli-Bless, C., 1999. [Quality assurance in nursing: self evaluation and peer review of nursing standards. Review of 2 years' experience]. Pflege 12, 187–193.

Schneider, J., Verta, M., Ryan, E., Corcoran, J., DaRosa, D., 2008. Patient assessment and management examination: lack of correlation between faculty assessment and resident self-assessment. Am. J. Surg. 195, 16–19.

Schroeder, A., Minocha, S., Schneider, C., 2010. The strengths, weaknesses, opportunities and threats of using social software in higher and further education teaching and learning. Journal of Computer Assisted Learning 26, 159–174.

Schuwirth, L., Van Der Vleuten, C., 2004. Merging views on assessment. Medical Education 38, 1208–1210.

Schwiebert, L., Davis, A., 1995. Impact of a required third-year family medicine clerkship on student self-assessment of cognitive and procedural skills. Teaching and Learning in Medicine 7, 37–42.

Scottish Government, 2008. Aspiring To Excellence - Scottish Government Consultation on Professor Sir John Tooke's Recommendations. http://www.scotland.gov.uk/Publications/2008/01/07144119/0 [accessed 24/9/11]

Seidenberg, M., Haltiner, A., Taylor, M.A., Hermann, B., Wyler, A., 1994. Development and validation of a Multiple Ability Self-Report Questionnaire. J Clin Exp Neuropsychol 16, 93–104.

Sharp, L., Wang, R., Lipsky, M., 2003. Perception of competency to perform procedures and future practice intent: a national survey of family practice residents. Acad Med 78, 926–932.

Shaughnessy, A., Slawson, D., 1999. Are we providing doctors with the training and tools for lifelong learning? BMJ 319, 1280–1280.

Sidhu, R., Vikis, E., Cheifetz, R., Phang, T., 2006. Self-assessment during a 2-day laparoscopic colectomy course: can surgeons judge how well they are learning new skills? Am. J. Surg. 191, 677–681.

Silver, I., Campbell, C., Marlow, B., Sargeant, J., 2008. Self-assessment and continuing professional development: the Canadian perspective. J Contin Educ Health Prof 28, 25–31.

Smith, G., Pell, J., 2003. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. BMJ 327, 1459–1461.

Smith, K., Tillema, H., 2001. Long-Term Influences of Portfolios on Professional

Development. Scandinavian Journal of Educational Research 45, 183–202.

Snadden, D., Thomas, M., 1998a. The use of portfolio learning in medical education. Medical Teacher 20, 192–199.

Snadden, D., Thomas, M., 1998b. Portfolio learning in general practice vocational training--does it work? Med Educ 32, 401–406.

Snadden, D., Thomas, M., Griffin, E., Hudson, H., 1996. Portfolio-based learning and general practice vocational training. Med Educ 30, 148–152.

Sobral, D., 2000. An appraisal of medical students' reflection-in-learning. Med Educ 34, 182–187.

Sobral, D., 2004. Medical students' self-appraisal of first-year learning outcomes: use of the course valuing inventory. Med Teach 26, 234–238.

Sommers, P., Muller, J., Ozer, E., Chu, P., 2001. Perceived self-efficacy for performing key physician-faculty functions--a baseline assessment of participants in a one-year faculty development program. Acad Med 76, S71–73.

Stewart, J., O'Halloran, C., Barton, J., Singleton, S., Harrigan, P., Spencer, J., 2000. Clarifying the concepts of confidence and competence to produce appropriate self-evaluation measurement scales. Med Educ 34, 903–909.

Story, A., 1998. Self-Esteem and Memory for Favorable and Unfavorable Personality Feedback. Pers Soc Psychol Bull 24, 51–64.

Strivens, J., Baume, D., Grant, S., Owen, C., Ward, R., Nicol, D., 2009. The Role of e-Portfolios in Formative and Summative Assessment: Report of the JISC-funded Study. Unpublished study, Centre for Recording Achievement for JISC. Retrieved on November 1, 2009.

Sullivan, M., Hitchcock, M., Dunnington, G., 1999. Peer and self assessment during problem-based tutorials. Am. J. Surg. 177, 266–269.

Swallow, V., Clarke, C., Iles, S., Harden, J., 2006. Work based, lifelong learning through professional portfolios: Challenge or reward? Pharmacy Education 6, 77–89.

Sweat-Guy, R., Buzzetto-More, N., 2007. A Comparative Analysis of Common E-Portfolio Features and Available Platforms. Issues in Informing Science & Information Technology 4, 327.

Teunissen, P., Scheele, F., Scherpbier, A., van der Vleuten, C., Boor, K., van Luijk, S., van Diemen-Steenvoorde, J., 2007. How residents learn: qualitative evidence for the pivotal role of clinical activities. Med Educ 41, 763–770.

Tillema, H., Smith, K., 2007. Portfolio Appraisal: In Search of Criteria. Teaching and

Teacher Education: An International Journal of Research and Studies 23, 442–456.

Tiwari, A., Tang, C., 2003. From process to outcome: the effect of portfolio assessment on student learning. Nurse Educ Today 23, 269–277.

Tochel, C., Beggs, K., Haig, A., Roberts, J., Scott, H., Walker, K., Watson, M., 2011. Use of web based systems to support postgraduate medical education. Postgrad Med J 87, 800–806.

Tochel, C., Haig, A., Hesketh, A., Cadzow, A., Beggs, K., Colthart, I., Peacock, H., 2009. The effectiveness of portfolios for post-graduate assessment and education: BEME Guide No 12. Med Teach 31, 299–318.

Tooke, J., 2008. Final Report of the Independent Inquiry into Modernising Medical Careers. http://www.mmcinquiry.org.uk/draft.htm [accessed 13/1/13]

Tosh, D., Light, T., Fleming, K., Haywood, J., 2005. Engagement with Electronic Portfolios: Challenges from the Student Perspective. Canadian Journal of Learning and Technology / La revue canadienne de l'apprentissage et de la technologie 31.

Tousignant, M., DesMarchais, J., 2002. Accuracy of student self-assessment ability compared to their own performance in a problem-based learning medical program: a correlation study. Adv Health Sci Educ Theory Pract 7, 19–27.

Tracey, J., Arroll, B., Richmond, D., Barham, P., 1997. The validity of general practitioners' self assessment of knowledge: cross sectional study. BMJ 315, 1426–1428.

Trudel, J., Bordage, G., Downing, S., 2008. Reliability and validity of key feature cases for the self-assessment of colon and rectal surgeons. Ann. Surg. 248, 252–258.

Turner, N., van de Leemput, A., Draaisma, J., Oosterveld, P., ten Cate, O., 2008. Validity of the visual analogue scale as an instrument to measure self-efficacy in resuscitation skills. Med Educ 42, 503–511.

UKFPO, Curriculum and Assessment. http://www.foundationprogramme.nhs.uk/pages/home/training-and-assessment [accessed 8/1/12]

Vallis, J., 2008. Personal Digital Assistants for Clinical Teams: Lessons Learned from Two Pilot Studies in a Scottish Teaching Hospital. The International Journal of Technology, Knowledge and Society 4, 1–10.

Van der Vleuten, C., 1995. Evidence-based education? Advances in Physiology Education 269, S3.

Van Tartwijk, J., Driessen, E., 2009. Portfolios for assessment and learning: AMEE Guide no. 45. Medical teacher 31, 790–801.

Van Wesel, M., 2008. The influence of portfolio media on student perceptions and learning outcomes. Presented at the Student Mobility and ICT, Maastricht.

Vecchioli, A., Ferro-Luzzi, M., Campioni, P., n.d. Assessment and self-assessment. Rays 15, 101.

Violato, C., Lockyer, J., 2006. Self and peer assessment of pediatricians, psychiatrists and medicine specialists: implications for self-directed learning. Adv Health Sci Educ Theory Pract 11, 235–244.

Wakley, G., 2000. Sexual health in the primary care consultation: Using self-rating as an aid to identifying training needs for General Practitioners. Sexual and Relationship Therapy 15, 171–181.

Ward, M., Gruppen, L., Regehr, G., 2002. Measuring self-assessment: current state of the art. Adv Health Sci Educ Theory Pract 7, 63–80.

Ward, M., MacRae, H., Schlachta, C., Mamazza, J., Poulin, E., Reznick, R., Regehr, G., 2003. Resident self-assessment of operative performance. Am. J. Surg. 185, 521–524.

Webb, T., Aprahamian, C., Weigelt, J., Brasel, K., 2006. The Surgical Learning and Instructional Portfolio (SLIP) as a self-assessment educational tool demonstrating practice-based learning. Curr Surg 63, 444–447.

Weiss, P., Koller, C., Hess, L., Wasser, T., 2005. How do medical student self-assessments compare with their final clerkship grades? Med Teach 27, 445–449.

Westberg, J., Jason, H., 1994. Fostering learners' reflection and self-assessment. Fam Med 26, 278–282.

Wetherell, J., Mullins, G., Hirsch, R., 1999. Self-assessment in a problem-based learning curriculum in dentistry. Eur J Dent Educ 3, 97–105.

Whitehouse, A., Hassell, A., Bullock, A., Wood, L., Wall, D., 2007. 360 degree assessment (multisource feedback) of UK trainee doctors: Field testing of team assessment of behaviours (TAB). Medical Teacher 29, 171–176.

Wikipedia, 2013. Web 2.0. Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Web_2.0&oldid=532416015 [accessed 1/10/13]

Wilkerson, L., Lee, M., Hodgson, C., 2002. Evaluating curricular effects on medical students' knowledge and self-perceived skills in cancer prevention. Acad Med 77, S51–53.

Williams, M., Jordan, K., 2007. The nursing professional portfolio: a pathway to career development. J Nurses Staff Dev 23, 125–131.

Windish, D., Knight, A., Wright, S., 2004. Clinician-teachers' Self-assessments Versus Learners' Perceptions. J Gen Intern Med 19, 554–557.

Woods, R., McCarthy, R., Barry, M., Mahon, B., 2004. Diagnosing smallpox: would you know it if you saw it? Biosecur Bioterror 2, 157–163.

Woolliscroft, J., TenHaken, J., Smith, J., Calhoun, J., 1993. Medical students' clinical self-assessments: comparisons with external measures of performance and the students' self-assessments of overall performance and effort. Acad Med 68, 285–294.

Young, J., 2012. "Badges" Earned Online Pose Challenge to Traditional College Diplomas. Chronicle of Higher Education.

Young, J., Glasziou, P., Ward, J., 2002. General practitioners' self ratings of skills in evidence based medicine: validation study. BMJ 324, 950 –951.

Zick, A., Granieri, M., Makoul, G., 2007. First-year medical students' assessment of their own communication skills: a video-based, open-ended approach. Patient Educ Couns 68, 161–166.

Zijlstra-Shaw, S., Kropmans, T., Tams, J., 2005. Assessment of professional behaviour--a comparison of self-assessment by first year dental students and assessment by staff. Br Dent J 198, 165–171.

Zonia, S., Stommel, M., 2000. Interns' self-evaluations compared with their faculty's evaluations. Acad Med 75, 742.